# Welcome to the course

[MUSIC]

Hi, my name is Nicolas Glady.

I'm a professor at ESSEC Business School.

I'm in charge of the Accenture Strategic Business Analytics Chair, and

I'm the Director of our Center for Digital Business.

I'll be your guide during your first steps as a strategic business analytics expert.

With this MOOC,

we'll have a first overview on strategic business analytics topics.

We'll discuss a wide variety of applications of business analytics.

From marketing to supply chain or credit scoring,

and HR analytics, we'll cover many different data analytics techniques.

Each time, explaining how to be relevant for your business.

We'll pay special attention to how you can produce convincing,

actionable and efficient insights.

We also present you with different data analytics too to apply

to different types of issues.

By doing so,

we'll help you develop four sets of skills needed to leverage value from data.

Analytics, IT, Business and Communication.

By the end of this MOOC,

you should be able to approach a business issue using analytics by

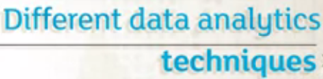
1, qualifying the issue at hand in quantitative terms and

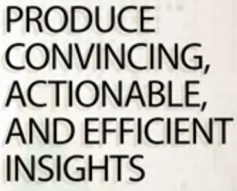
2, conducting relevant data analyses and

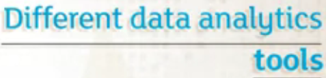
3, presenting your conclusions and recommendations in a business oriented,

actionable, and efficient way.

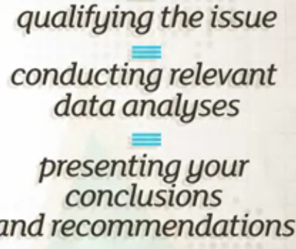












# Becoming a Business Analytics expert

[MUSIC]

With this course, you'll become an expert on strategic business analytics.

As we'll explain later,

we'll help you develop four sets of skills needed to leverage value from data.

Analytics, IT, Business, and Communication.

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

We'll discuss a wide variety of examples from the business world.

Different functions such as marketing or supply chain, and

different sectors such as the telecommunications sector,

the retail sector, the public sector, and so on.

In order to analyze the related data.

We'll present different techniques, use the strategic business analytics.

How to find groups of similar and dissimilar observations.

How to find relationships between causes and consequences.

How to make forecasts, and how to make all those techniques relevant for

your business.

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

Actually, we will pay special attention to the business aspect of those questions.

Play video starting at :1:9 and follow transcript1:09

What is the actual problem we want to solve?

And how can we make our conclusions really convincing, actionable, and efficient?

Play video starting at :1:19 and follow transcript1:19

At the end of each module, there will be a recital of the different techniques for

helping you to wrap up the tools and techniques presented during this module.

Play video starting at :1:31 and follow transcript1:31

There will be then a quiz to complete.

For the last module, we will require you to produce a business level output

of your analysis that will be evaluated by the other participants.



# Why? It is all about value not data

[MUSIC]

This MOOC is called Strategic Business Analytics.

As you notice we do not use the word data in our title.

It's because all the questions we'll answer in this module will be driven by

and for business, and be focused on strategic issues.

Data is a means, neither an end nor even a beginning.

Consequently, we'll mainly focus on the questions and

solutions that data analytics can provide.

We'll neither focus on how data analytics tools work,

nor even what are data analytics tools,

making a kind of list of all the existing techniques.

But we can instead cover different business issues that can be solved

leveraging data, and we provide some useful tools that apply in this context.

We'll then focus on how to present and

articulate those results in a meaningful way.

Too often when dealing with data

we see it as a means to understand a situation better, and we stop there.

It's actually not the end of it.

The purpose of data analysis is to understand the problem

we want to solve in order to define business actions.

So the real focus of this course will be on how data could help to

make actionable recommendations.

Data is very often referred to as the new eye.

We'll approach it as an asset which can be used to leverage value.

You don't need to be an engineer to understand why oil is valuable, or

how it can be used to improve your business.

Some people may be interested in the chemistry or physics of oil, but

that's not really what business leaders focus on.

In contrast, we want to focus on the managerial aspects of it.

What are the tools we need to analyze the data?

What does the output of those tools mean for business?

And we'll leave the technical parts to other trainings,

such as those specialized in art, or statistics.

All the following videos will be articulated around this approach,

as it is of course for managers and

leaders that are interested in using data to make better decisions, but are not

necessarily willing to develop their statistical or computer science skills.

Play video starting at :2:38 and follow transcript2:38

What is the business issue that data can solve?

What are the relevant conclusions of data analysis?

How can we make those conclusions

actionable to improve our business efficiency?

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

You can approach this training as an introduction to analytics

from the business perspective.

That will present you different applications and techniques.

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

Techniques that you can then understand in details thanks to other trainings.

Or, if you're in the opposite situation and if you already are a data scientist,

statistician, or the like, as a way, to understand what are the business

applications of analytics techniques and tools that you already know, but

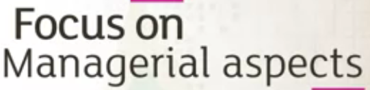
that you could never practice in a business context.

Play video starting at :3:30 and follow transcript3:30

Our program is not aiming at developing your data science skills,

even if we expect you to be able to use some statistical and data analysis tools.

We want to train you to become data-driven managers.



# How to leverage data for value - from data to insight

[MUSIC]

So the focus of this course will be to create value by making actionable

recommendations based on data analysis to solve current business issues.

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

This means two things: first we'll not focus on technicalities.

There are a lot of data analytics libraries that can be used

"out of the box" nowadays.

Play video starting at ::32 and follow transcript0:32

And there is little added value for

the manager to understand all the little gory details.

One can use the default settings of data analytics tools and

make actionable recommendations.

But that means that two: one needs to be able

to understand what are the conclusions of the obtained results.

Play video starting at ::53 and follow transcript0:53

What does a significant factor mean?

If we see a pattern in data, what can we do with it?

Those are the type of questions we'll be focusing on,

which means that we expect the participants to have a sufficient

level in R programming for them to be able to produce meaningful results.

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

We won't detail too much how the R libraries work or

what can be done to tweak the parameters.

Instead, we'll discuss which libraries can be used and

how to interpret the resulting output.

Play video starting at :1:27 and follow transcript1:27

Note, however, that you will have the opportunity to rehearse and

practice the different examples presented during the videos

thanks to the recitals at the end of each module.

Play video starting at :1:39 and follow transcript1:39

Those recitals will go over the R code in detail.

Then after analyzing the data, if we want our recommendations to be impactful

they need to be adopted by the different stakeholders of the focal company or

organization.

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

Hence, we'll also discuss how to present data analytics results in a convincing and

impactful way.

This is too often overlooked...

Play video starting at :2:4 and follow transcript2:04

while if you want the many hours invested in your analysis to be fruitful,

you want your conclusions to be understood

And accepted.

Play video starting at :2:13 and follow transcript2:13

This relies on effective communication.

That's why this course is articulated around four core skills development:

One: analytics in the strict sense.

We want you to be able to produce relevant statistics.

Two: to do so you need IT skills. This course will focus on R,

but our program goes way beyond the usage of a mere programming language.

Play video starting at :2:37 and follow transcript2:37

As explained before, what we want is really to focus on three: the business.

We'll cover many functional domains

and relate our analysis to business issues and solutions.

But for those solutions to be accepted by the different stakeholders

you need: Four, communication skills.

Play video starting at :2:58 and follow transcript2:58

You need to be able to express your complex ideas in simple terms

and to adopt the right format to the right audience.

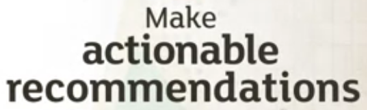
Those are actually consulting skills.

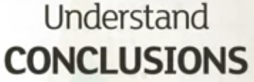
Later in this specialization, you'll actually have your opportunity to be

exposed to worldwide-level expert consultants from Accenture.

And we want you to be able to start your journey as a world-class

business analytics consultant right now.







# Dataset for practice quiz

Dataset for the following practice quizs

[PASTAPURCHASE\_EDITED](https://d3c33hcgiwev3.cloudfront.net/_eba2c079135882131db3690701bc9c97_PASTAPURCHASE_EDITED.csv?Expires=1700092800&Signature=Dg-r~slcNlfqFdm38yRWYJnejnyoSVhdmYgozH~3vHnozts8fY-uhZQW7~JOzaLM7BJ95Y31VdxoWOoL13xGEc~k~bLU3gTNv2D9HgZXk2hzW~DX~FR0NaB81VpH2O92NA66azCFc1Lph9wCVCEHMcpKt6IlA-l7KyHUxfsqPvY_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[CSV File](https://d3c33hcgiwev3.cloudfront.net/_eba2c079135882131db3690701bc9c97_PASTAPURCHASE_EDITED.csv?Expires=1700092800&Signature=Dg-r~slcNlfqFdm38yRWYJnejnyoSVhdmYgozH~3vHnozts8fY-uhZQW7~JOzaLM7BJ95Y31VdxoWOoL13xGEc~k~bLU3gTNv2D9HgZXk2hzW~DX~FR0NaB81VpH2O92NA66azCFc1Lph9wCVCEHMcpKt6IlA-l7KyHUxfsqPvY_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

<https://github.com/nicolasfguillaume/Strategic-Business-Analytics-with-R/blob/master/module1.md>

# Introduction: what’s the point of finding groups within data?

In this module, we'll first explain how identifying groups of

observations within data allows to improve your business efficiency.

Then we'll explain how to keep those groups in a business oriented and

actionable way.

Play video starting at ::23 and follow transcript0:23

We'll take on key examples to illustrate the different concepts we'll cover.

Finally, we invite you to replicate those examples

thanks to the recital provided at the end of the module.

It's important to understand that we won't be focusing on the technical or

statistical details of data analytics tools we'll be using during this module.

Play video starting at ::44 and follow transcript0:44

We'll use those tools as a means to produce results that will analyze and

try to make action.

When you're using a car, you don't need to understand how the engine works

to drive it to your destination.

It's a means to achieve your objective to go from a point A to a point B.

Play video starting at :1:3 and follow transcript1:03

Our approach will be exactly similar.

Play video starting at :1:6 and follow transcript1:06

We'll use some tools, just giving the intuition of what's their purpose, but

we won't describe all the parameters of the function, or

cover all the hypotheses needed for the statistical model to work.

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

Instead we'll focus on one: how to come quickly to relevant statistics and

two: how to interpret those results and make all conclusions

actionable in the context of the business case we are trying to solve.

What does finding groups within data mean?

Finding groups within data and distinguishing observation in

your data set that are similar from those that are different.

Why do we want to find groups within data?

The objective is to find the right balance between similarities and differences.

On the one hand we want to treat similar cases in a similar way

to benefit from economies of scale.

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

But on the other hand, we want to treat different cases

in a different way to improve your actions' effectiveness.

Play video starting at :2:14 and follow transcript2:14

And when you are trying to find the right balance between economies of scale

Play video starting at :2:20 and follow transcript2:20

and customized, effective actions,

what you're actually trying to do is to improve your business efficiency.

Play video starting at :2:28 and follow transcript2:28

That's exactly how you need to see it.

Finding groups within data may look like a technical exercise at first, but

its final purpose is to be instrumental in maximizing your business efficiency and

helping you solving problems that are fundamentally business ones.

For instance, imagine that your supply and logistics manager

was willing to organize his company's stock management more efficiently.

Or imagine that you are an HR manager and want to see whether there are patterns

in the behavior of the employees leaving your company.

Play video starting at :3:7 and follow transcript3:07

A third example could be that you are a product manager,

willing to understand better how to serve your customers.

Play video starting at :3:17 and follow transcript3:17

In all those cases, finding groups within data will allow you

to allocate your effort more efficiently.

One, by leveraging synergies between cases that are similar and

two, allocating specific case houses to different cases if needed in

order to maximize your effectiveness.

# Basic clustering using ad-hoc techniques: the example of product management

Let's start with the first example.

Imagine that you're working as a supply and logistics manager, and

want to organize your supply chain more efficiently.

Play video starting at ::19 and follow transcript0:19

You realize that you have some products,

with sales that are more variable than others.

It's impacting your ability to deliver the right goods at the right time.

On the one hand, you could decide to maintain huge stocks for

all your products, and you would ensure product availability by doing so.

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

But it would be really expensive, to implement in practice.

On the other hand you could wait up to the last minute,

before deciding on the production.

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

Which would be cheaper, but

your ability to deliver the product on time would be negatively impacted.

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

A simplistic approach to this problem, but similar to what's used in industry,

would be to analyze the question along two dimensions.

You could plot the average daily sales of the stock keeping unit,

the SKU, as a function of the volatility of daily sales of this SKU.

The volatility of something may be measured with a coefficient variation.

For instance, in this case the coefficient variation,

is the standard deviation of sales, divided by the mean sales for this SKU.

We can then identify groups of SKU that are similar to each other visually.

Play video starting at :1:38 and follow transcript1:38

For instance, we can split both the variability and

the average sales dimension in two, and

group the two bottom cases that is SKUs with smaller average sales together.

We then have three cases.

One, SKUs with high sales and low variability,

let's call them "horses" because those products are strong and reliable.

Two, a scale used with high sales and high variability.

Let's call them "white bulls", because they are strong, but difficult to control.

And three, let's consider all of the other ones together, low sales, and

call them "crickets" because they are small, but can jump unexpectedly.

We'll then manage our supply chain differently for

those three cases, in order to ensure maximum efficiency.

Horses supply chain will be organized as what we call, "make to stock".

They'll be made quickly available for sales by forecasting the demands,

and preparing the production in advance.

It may seem expensive to keep everything in stock, but

the benefits cover largely the costs.

Because sales are expected to be high, and the risk of inaccurate forecast is low,

since the coefficient of variation is small in this case.

Crickets supply chain, will be "made to order".

We'll wait to have an order before starting the production process.

Since the sales are small, in any case,

it's not really efficient to prepare production too long in advance.

So we want to reduce the risks by producing the goods,

only if the order is made by the end of the chain.

White bulls will be treated on a case by case basis.

It's okay because there are only a few of them.

Things may difficult to anticipate, but their return may be high.

So let's be pragmatic and access the situation in each case specifically.

In practice, you can multiply the splits by dimensions.

For instance if you do a three by three analysis,

you would have what is called an abc- xyz analysis.

And you may have a final segmentation of SKUs.

And a resulting design of your supply chain that is more detailed.

Note how clear our result is.

We limited our analysis to only two dimensions here.

But it makes the results very easy to understand.

And the fact that we decide to use images and names for the different groups of

products, will ensure that audience will remember our conclusions.

It will foster the adoption of the message we want to convey.

When you can, always try to summarize your conclusion in a two by two matrix.

Or at least with two axes, as we did here.

And always name your segments in a meaningful, self explanatory and

memorable way.

Play video starting at :4:39 and follow transcript4:39

Even if it's not always easy to do, a picture is worth a thousand words.

During the different examples we'll cover, we'll always try to come up with a visual

representation that makes all the results more meaningful and convincing.

Note that all those examples are usable for other applications,

as we provide an R script during the recital at the end of this module

that is easily adaptable to other contexts and other datasets.

# Identifying groups within data: what's the intuition behind clustering? The example of HR Analytics

Now in practice there are only a few situations where you can define your goals

visually because in practice, you very often have more than two dimensions of

variables that you may consider for your clustering.

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

So we need to rely on a more rigorous approach to create groups and

identify patterns within the data.

Play video starting at ::31 and follow transcript0:31

An approach often used in practice is the hierarchical clustering technique.

The recital at the end of this module will give you an opportunity to

learn how to perform hierarchical clustering in R yourself

in particular, using the hclust function.

Note that you can replicate all the examples that are presented in this video

thanks to the scripts and datasets that are provided with this MOOC, and

discussed during the recitals.

So let's apply hierarchical clustering to the previous example.

I said that what's important is to find the right balance between treating similar

cases similarly, and different cases specifically.

This is exactly the same from a clustering perspective as if to say that we

want to maximize the similarity within the clusters, and

maximize the dissimilarity between clusters.

In statistics,

the similarity is very often measured with the distance between observations.

We may, for instance,

take the Euclidean distance, that is the distance you probably know from geometry.

What the hclust function does is - for

a given number of groups - to identify the optimal clusters in that sense.

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

Let's go back to the SKU example.

And - as you can test yourself during the recital - we can have three groups that

maximize the within-cluster proximity, and maximizes the inter-cluster distance.

Play video starting at :2:11 and follow transcript2:11

And what's neat is that we'll find exactly the same groups as we did before but

now with a more vigorous approach.

Play video starting at :2:18 and follow transcript2:18

We now have a tool that is statistically more robust but also more flexible.

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

First, as in practice the question of "how many groups should we have?"

is very often crucial.

We can now select the number of groups as a parameter of hclust.

Second, we can also include as many dimensions in the analysis as we want.

Not only 2 or 3 but even 10 or 20 if needed.

Should we have 20 dimensions,

finding groups visually would actually be impossible.

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

By the way, about the number of groups to define, some people

may claim that you should use technical criteria such as the dendrogram to

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

decide how many clusters you should have in practice.

In contrast, in this training, we claim that you should mostly rely

on common sense and business criteria to decide on the number of clusters,

Play video starting at :3:20 and follow transcript3:20

while ensuring that clusters you pick are still statistically relevant, obviously!

Play video starting at :3:26 and follow transcript3:26

In the supply chain example, we decided to keep only three groups

because the splits along those axes are easy to communicate.

Play video starting at :3:35 and follow transcript3:35

It's a common sense criteria.

Later we may be constrained by other factors

such as the capacity of your workforce or the current organization of your company.

Play video starting at :3:48 and follow transcript3:48

We'll discuss such examples later.

But at the end of the day, you need to understand that business analytics

aims at finding the right balance between business relevance,

technical accuracy, and common sense.

Thanks to the hclust function, which provides a hierarchical clustering tool,

Play video starting at :4:9 and follow transcript4:09

we can now approach more complex examples

where we have more than two dimensions to study.

In practice, you will very often have to select only part of the variables that you

can use to make groups because all the data is not usable.

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

I'll skip the discussions on data sampling, technical variable selections,

or missing data.

And I'll focus on the core of our approach: can we select variables that

are more relevant from a business actionable perspective?

And for that, let's imagine that you're an HR manager of a big consulting company.

and that you're concerned by the number of employees leaving the firm.

As an HR manager,

you want to retain your best employees within the company but you cannot

follow-up with each one of them too often as that would be very time consuming.

Play video starting at :5:4 and follow transcript5:04

Instead, you may rely on an HR analytic solution to understand what are the most

frequent situations explaining why an employee decides to leave.

Now let's imagine that this company has collected a bunch of data

during exit interviews.

Such data would be information like the satisfaction of the employee about

the company, the last project evaluation, the number of completed

projects within the last 12 months, the average amount of hours worked per month,

the time spent in the company, and whether he or she had a baby within the last year.

We first need to make the variables comparable.

For instance,

we cannot really compare the age of an employee with her satisfaction level.

What we can typically do in practice is to standardize the variables.

To do this,

we first subtract to each value from each variable the average for this variable.

This is called mean-centering.

And then we divide by the standard deviation.

In doing so we make the distance along variables comparable.

We explain in the recital how to make this standardization in

practice with the scale function.

Now we just have to apply the hierarchical clustering.

Play video starting at :6:26 and follow transcript6:26

I won't go through the technical details because I've already provided

all the intuitions needed, and

we explain during the recital how to implement it in R.

We'll discuss later how we can decide on the number of segments in a more

systematic way. But let's imagine that we decide to have 4 segments.

We run hclust and

after some transformations we obtain the following output.

How can we interpret the results?

Play video starting at :6:54 and follow transcript6:54

We need to understand how the segments are different.

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

We could do an ANOVA, an analysis of variance to be systematic. But

let's just try to visualize the most noticeable differences with \*colors.

Play video starting at :7:9 and follow transcript7:09

The first segment didn't do many projects on average and was underutilized.

Play video starting at :7:16 and follow transcript7:16

It's also a segment where employees have been in the company for

a shorter time than the average.

Play video starting at :7:23 and follow transcript7:23

Let's call them the "low performance segment".

The second segment has a low level of satisfaction.

The good evaluations worked on lot of project,

has a good utilization rate, and has been in the company for a long time.

Play video starting at :7:39 and follow transcript7:39

We see that there are those who works the most, but also, they're less satisfied.

The two events could be related, so let's call those employees "the burned" ones.

The third segment is very similar to the second segment, with good evaluations and

high utilization. But in contrast, they are very satisfied.

Let's call them "the High Potential".

And finally, the last segment doesn't have any distinctive characteristics.

Let's call them the "Misc" segment.

First, we need to say something about the newborn variable.

In practice, on the one hand, we could consider this variable as exogenous,

which means that we can measure it, but

except if you're in a country that limits the number of children a family can have,

we cannot do anything about it in practice.

Play video starting at :8:31 and follow transcript8:31

So this variable will never be really actionable in any case.

From this perspective,

the variable should not have been included in the data set in the first place.

On the other hand, the firm may consider developing parents' benefits,

such as paternity or maternity leave, or on-site childcare.

So in practice,

we see that the question of which variable should be taken into account in

the analysis is actually related to its actionability and it's relevance.

It's a business discussion and not a technical one.

Play video starting at :9:6 and follow transcript9:06

Second, we should note that the level of satisfaction is a consequence of

everything else.

We cannot really act on it directly. So

it has to be seen as a consequence and not as the driver of managerial impact.

And we should focus on what is actionable in practice.

Play video starting at :9:25 and follow transcript9:25

The number of projects done or the utilization, for

instance, can directly be acted upon by the management.

We can staff more or

less a certain person if we see that the situation needs to be changed.

Play video starting at :9:40 and follow transcript9:40

The conclusions are then straightforward.

For the low performance, whether they leave the company themselves or

are fired, it may not be a priority to try to retain them.

Play video starting at :9:53 and follow transcript9:53

In contrast, we should really do something quickly for the "burned out"!.

The manager should have anticipated the situation and

helped those employees to take a step a back proactively.

Now, for the High Potentials, the situation would be more complex.

Play video starting at :10:9 and follow transcript10:09

Those employees have probably been hired by the client.

It's probably that we couldn't make them a better offer.

We can always try to give them a raise, or promotion, or better projects,

but it's always difficult to retain those employees that are happy but

want to leave nonetheless.

Finally, we very often observe a "misc" segment

whatever the situation or the sector.

We cannot always explain everything.

And some events may be exogenous:

explained by events that are out of our control.

We may want to collect additional data, to try to understand this specific case

better but since it's a relatively small group, it's not really a priority either.

Notice that in all the cases, we start from the business issue,

sub-optimal supply chain design or employees leaving the company.

And it's the business issue that drives the type of variables

we are interested in.

Play video starting at :11:8 and follow transcript11:08

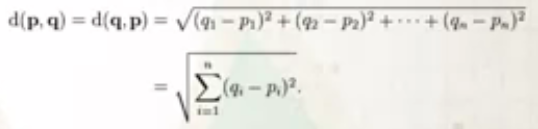
The same for how we design our analysis or our conclusions.

You have to understand that there is no such thing as on "optional clustering or

segmentation approach" in practice.

What matters is that you use the right clustering approach for

the business problem at hand, and that your conclusions are actionable.





# Introduction to Customer Segmentation

Let's finally take an example of customer segmentation because clustering

is very often used in marketing for identifying customer segments.

Play video starting at ::20 and follow transcript0:20

Marketing analytics is, as a matter of fact, probably one of the most,

if not the most, important fields in business analytics.

That's why we decided to include a specific marketing analytics MOOC in

this specialization as well.

So let's take a relatively simple example of customer segmentation.

Play video starting at ::40 and follow transcript0:40

In all the previous examples we've mainly discussed

how to identify different groups.

We didn't really assess whether all those groups were really different.

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

This is important because since we want to reduce the effort needed to act

effectively, it may be that in some cases some groups of observations, here some

customers, even if slightly different, would actually be treated similarly.

For instance, if you have two groups of customers that have different ages or

slightly different product usages, but that are at the end of the day served

in the same way, why should we consider them as different segments?

Well, we shouldn't.

Play video starting at :1:26 and follow transcript1:26

Let's take a concrete example in the telecommunications sector.

Play video starting at :1:30 and follow transcript1:30

Let's say we have a customer base for which we have different information.

Play video starting at :1:36 and follow transcript1:36

For each customer, we know the average monthly national call time.

The time spent in international calls, the number of text messages sent,

the data used, and the age of the customer.

We first need to normalize the variables again.

They don't have the same units.

So if we don't want to compare apples and

pears, we'll apply the R scale function again.

Let's start to test what would be the result for eight segments.

We can then compute the average value for each of those variables per segment.

Play video starting at :2:13 and follow transcript2:13

This table reports the average values per segment

provided by the Hclust function for the case we selected.

Play video starting at :2:22 and follow transcript2:22

We see for instance that the first segment

doesn't have a high level of usage but is older than the other ones.

Play video starting at :2:31 and follow transcript2:31

Let's call this segment the Silver.

Play video starting at :2:34 and follow transcript2:34

The second segment makes a lot of national and international calls,

but fewer text messages.

Let's call them the Pro.

Play video starting at :2:43 and follow transcript2:43

We can then identify the segments, young adults, 30-something,

40-something, teens, low usage, and experts.

You can look at the exact data with the recital and the R could provide it.

But now we notice that there are segments that, while statistically different,

would actually be targeted by similar offers.

For instance, the young adults and the teens clearly have similar usages,

with fewer calls but a lot of text messages and data.

Play video starting at :3:17 and follow transcript3:17

Similarly the thirty-something and

the forty-something are also very similar to each other.

So let's apply Occam's Razor and

let's run the hierarchical clustering again but with fewer segments.

Play video starting at :3:31 and follow transcript3:31

In practice,

we could run several tests but here we'll just focus on the case with five segments.

Play video starting at :3:38 and follow transcript3:38

We obtain another table, what do we see?

We can distinguish, one: a heavy user segment making a lot of calls,

using a lot of data and text messages.

Two: a young-adult segment using a lot of data and texts, but fewer calls.

Three: a Silver segment with a low usage overall, but more seasoned.

Play video starting at :4:2 and follow transcript4:02

Four: a pro segment with a lot of national and international calls, and

a lot of data.

Play video starting at :4:8 and follow transcript4:08

Five: finally, a light user segment.

We created groups of similar customers to solve them efficiently.

Play video starting at :4:17 and follow transcript4:17

An efficient marketing strategy will define for each segment, the bundle to

offer, at the right price, on the right place, with the right message.

This message will be a function of what the customer needs but

also their social demographics.

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

Here we only had five variables.

But in practice,

we can take as many variables as long as the data is available and is relevant.

Play video starting at :4:43 and follow transcript4:43

We can also try to find a more visual representation than a table.

You can use the many visualization libraries that are available in R.

Such as one providing a radar chart, for instance.

A radar chart is a neat visual representation

because it allows to compare multi-dimensional elements.

Each branch of the radar indicates the scaled value on

each dimension of the segmentation.

Play video starting at :5:11 and follow transcript5:11

We can also do the same by exporting a table into Excel and

creating the radar chart there.

Play video starting at :5:17 and follow transcript5:17

Actually, that's what I would do in practice because

I suppose you'll agree that R rarely produce

pixel perfect figures that can directly be shown to an external audience.

Play video starting at :5:29 and follow transcript5:29

At the end of the day, you want to make a presentation that is beautiful.

Play video starting at :5:34 and follow transcript5:34

It doesn't matter if it's done in R or in something else.

Actually, nobody cares which tool you use.

But if it's ugly, it won't give the impression to your audience that your work

is of high quality.

Play video starting at :5:47 and follow transcript5:47

So when you report your results, for instance, as you will have to do with

the last assignment of this MOOC, or the capstone project

Play video starting at :5:57 and follow transcript5:57

pay special attention to the presentation of your results and

make sure the visuals are relevant, clear, and beautiful.

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

We now can say that we've aimed at the best balance between cost and

effectiveness.

The cost-to-serve is as limited as possible,

because we reduced the number of segments ensuring economies of scale, and

we implement specific actions for each segment having specific needs.

That is, we ensure the effectiveness of our actions.

The right balance between costs and effectiveness makes our actions efficient.

Play video starting at :6:37 and follow transcript6:37

We see that we use IT and statistics in conjunction with business acumen

to converge on actionable and efficient conclusions and recommendations.

For instance, to solve the question of how many segments we should have,

we start from a reasonably large number of segments and

we try to reduce it based on business common sense.

How will it be implemented in practice?

And how to achieve efficiency?

We also pay special attention to how those conclusions and

recommendations are reported so that the resulting output is clear and

obvious even for a non-technical audience.

Play video starting at :7:19 and follow transcript7:19

You will now have the opportunity to practice yourself thanks

to the recital of the R script that will replicate

the different examples we have discussed in this module.

# Presentation of Pauline Glikman

Hi, there.

My name is Bowen Goodman, and I am a student here at ESSEC Business School.

Play video starting at ::6 and follow transcript0:06

I'm delighted to be your tutor for this recital.

Play video starting at ::10 and follow transcript0:10

The purpose of this recital and all the following is for

you to put in order the concepts learned during the lectures.

Play video starting at ::19 and follow transcript0:19

As professor Gladys says in class, all the analyses we perform in this course

aim at producing business insights, and allow for actions to be taken.

Play video starting at ::29 and follow transcript0:29

With this in mind, we will try to keep our goods simple, and produce visualizations

that can be shown used and understood by a business audience.

Play video starting at ::39 and follow transcript0:39

Ultimately, our goal is that these recitals give you toolkit needed to

start solving your own problems and writing analysis and

data sets that are relevant to you, your business or your organization.

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

Also note that we provide you with our script of this tutorial so

you can reuse it and adapt it to perform analysis on your own data sets.

While you may download the script and

just run the code, we encourage you to open an empty script and follow along

with the tutorial in order to make sure that you understand the code line by line.

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

If you need time, do not forget you can always post a video and

come back to it later once you have updating your code.

This week, you saw in class that clustering techniques can help us find

groups, patterns or segments within data.

Those clusters can provide us knowledge that we may be able to convert

into actions that will themselves lead to improved business efficiency.

In this recital,

we will cover all of the examples that you covered with Professor Glady in class.

Namely, the SKU example, the HR analytics example and the customer user example.

We will provide a line by line analysis of the code used to obtain

the outputs that you show in class as well as discover ways to present

results in a way that is meaningful to a business audience.

Play video starting at :2:13 and follow transcript2:13

Let's get started.

# Recital M2 - SKU example

In this MOOC, we will be using RStudio, which is a programming environment for R.

Let's go ahead and open the script for this module in RStudio.

Play video starting at ::15 and follow transcript0:15

The first example we saw in class is concerned with improving a supply chain.

We will explore our data set to identify different groups of products

that share similarities and that can be treated in a similar way.

The first thing that we want to do is to set our working directory to the folder

where we have downloaded the SKU data set.

The way you do this in RStudio is that you navigate here to the right folder.

And then you click on More and

Set as Working Directory once you are in the right folder.

Next, it's usually a good idea to clean up the memory of your current R session.

And you can do that by executing this line.

By the way, in order to execute a line, you can highlight it, and

then press Command+Enter on a Mac and Ctrl+Enter on a PC.

Let's go ahead and do that.

We are now ready to move on and load our data set.

Let's call the data set DATA.

And we'll use the read.table function in order to read the data set.

The read.table function lets us read flat files such as CSV files into R.

The first argument of the read.table function is

the name of the file that we want to load.

Don't forget to put it in quotes.

The following argument here is the header argument,

which tells R that our original file contains a header.

So we set it equal to true.

And then we have the sep argument which tells R

the type of separator that are contained in our original file.

Let's now run this line in order to load the data set.

Play video starting at :2:7 and follow transcript2:07

Once your data is loaded the first thing you should always do is to explore

your data set.

First, let's use the STR function and call it on data.

Play video starting at :2:21 and follow transcript2:21

Let's have a look at the output.

What the STR function tells us is that we

have 100 observations in this data set and two variables.

What are the variables?

Play video starting at :2:32 and follow transcript2:32

The ADS variable, which is short for

average daily source as you may remember from the lecture.

And the CV variable, which is short for coefficient of variation.

The STR function also tells us the type of variable that we have.

Here you can see int which is R's way of telling us

that the variable contains integers only.

And here, for the CB variable, you can see num, which is short for numeric and

which is R's way, again, to tell us that the variable contains decimal numbers.

The next step would be to have more information about our variables,

such as their average and some other statistics.

And to do so, we can use the summary function, which you can see here.

Play video starting at :3:25 and follow transcript3:25

Let's run the line.

Play video starting at :3:27 and follow transcript3:27

As you can see, what the summary function does is to give us,

for each variable, the minimum value, the first quartile, the median,

the mean, the third quartile, and the maximum value.

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

Since we only have two variables, it seems like a good idea to plug our data,

to see visually if we can already distinguish different groups of products

with similar coefficient variations and average daily sales.

To plot data in R, you use the plot function.

Play video starting at :3:58 and follow transcript3:58

Your first argument is the coordinates on the vertical axis.

So here, that is the coefficient of variation.

And your second argument is the coordinates on the horizontal axis,

which is here, the average daily sales.

With the main argument, you can add a title to your plot, and

similarly with the y lab and the x lab argument,

you can label your variable in order to make your plot much more understandable.

Let's now run this line of code.

[NOISE] At this point, you should obtain this plot

that you can see on the bottom right of my screen.

What can you say about this plot?

Do you notice different groups of products that are close to one another?

If yes, how many?

Hopefully you are able to identify three groups of products on the plot.

The first group contains the SKUs with a coefficient of variation below 0.2 and

average cells between 4 and about 14.

As explained during the videos, those are our horses.

Our second group contains SKUs with a coefficient

of variation around 0.35 and about 1.

And between six and about fourteen for the average daily sales,

and as you may remember from the lectures, these are our world balls products.

And the third, and last, group contains SKUs with

a coefficient of variation between 0.2 and about 0.7 and

an average daily sales between 0 and about 3.

These, as you may remember, are the crickets.

Play video starting at :5:53 and follow transcript5:53

Let's draw some lines and

add some text to our plot in order to make it more explicit.

To add a vertical line in R,

you can use the abline function and you pass it two arguments.

The first one is v.

Since we identified that our first group, as coefficient of

variation b low point two, we set v equal to 0.2.

And our second argument, here, is the col argument, which controls the color,

so we set it equal to red in quotes.

Play video starting at :6:26 and follow transcript6:26

Let's run the line.

Now let's add a horizontal line by using the abline function again, but

this time we pass it the h argument because it's a horizontal line.

And since we identified our first group to have average daily sales over four,

let's set it equal to four and again, we'll set the color to be red.

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

To make our plot more visible, let's set a label to where our first group is.

Our first argument is the x value of where we want our text to start,

the second is the y value of where we want our text to start, and

the third argument is the text that we actually want to insert.

And then now you're familiar with the col argument, which sets the color.

And that we'll set here equal to red.

Play video starting at :7:21 and follow transcript7:21

Now you can do the same thing for

the wild bulls and the crickets by running these lines.

Play video starting at :7:28 and follow transcript7:28

Now if you try to move your text and lines around a bit, you will find that producing

nice visualizations in art can be time consuming and

is generally not the most efficient way to produce a business visualization.

Using a presentation tool, such as Microsoft's PowerPoint or Apple's Keynote,

and adding pictures and simple one liners will most often help you convey

the points that you're trying to make in a business-oriented manner.

Play video starting at :7:57 and follow transcript7:57

Feel free to use any presentation tool that you may be used to and

to post the visualizations you build in the firm, so

that we can all learn from each other's creativity.

So by simply plotting our data,

we were able to find groups of products that are similar to each other.

Let's now try to use hierarchical clustering methods learned in class, and

see if we can obtain a similar result.

First, in order to make sure that we do not alter our original data set,

let's make a copy of it and call it test data.

The way you do this is that you just set test data equal to data.

As we saw in class,

in order to make the variables comparable to one another, we need to normalize them.

To do so in R, you can use the scale function.

Let's use it.

We are now ready to compute the distances between our data points

which we will sure indeed.

Play video starting at :8:53 and follow transcript8:53

In R, we can do so by using the dist function which takes two arguments.

The first one is the data set that you want to compute the distances off,

and the second one is the method that you want to use in order to compute distances.

In this case, we're going to use the euclidean method, so

you set the method augment to euclidean and do not forget the quotes.

Play video starting at :9:21 and follow transcript9:21

Now that we have computed distances, we can finally use the hclust

function in order to perform the actual hierarchical cluster.

Let's create a new variable called hcword which will store the result of

the col of the hclust function.

The hclust function takes two arguments.

The first one is the result of our dist function, so d.

And the second argument is the clustering method that we are going to use.

In this case, we set it equal to word.d.

The question of the clustering method that we wanted to select, is a technical one.

But if you want more information about the hclust function, you can always type

?hclust in the control.

Play video starting at :10:14 and follow transcript10:14

And you obtained documentation on the package, or

you can always search for more on the web.

Let's now run our hclust function.

Play video starting at :10:26 and follow transcript10:26

We are now ready to assign all our observations to our groups and

we can do this using the qtree function.

Play video starting at :10:37 and follow transcript10:37

It takes two arguments, the output of the hierarchical

clustering in AC word and k equal to the number of clusters that we want to create.

As we said in the lectures, we choose to have three different clusters.

What we're doing in this line is that we're creating a new variable to

our original data set.

Which was data and not test data.

And to do so, we just write data$, the name of the new variable here,

groups, And here, instead of equal, I've used these two characters, but

I only did it so that you guys are aware that we can use this instead of equal.

What we're doing here is that we're creating a new column to

the original data set.

And the one that's called data, not test data.

And what this common indicates is for

each SKU, in which group the SKU is.

According to the hclust function.

Now that we have our groups, we can finally plug them.

The lattice package provides a lot of functions

that you should explore to make nicer plots.

In order to install a package in R, what you do is that you use

the install.packages function and you insert the name of the package in quotes.

Let's run it.

Play video starting at :12:3 and follow transcript12:03

Then, you need to load the package.

Which you can do using the library function,

and inserting the name of the package without quotes this time.

The plot that I made with the lattice function is the fill in one.

Play video starting at :12:21 and follow transcript12:21

Essentially what it does, is that it plots each group in a different color.

As you can see, we can find the result that you show

with Professor Glady in the lectures, which is very,

very similar to what we already saw by just simply plotting the data onto axis.

This wraps up our first example, and I will see you in the next video.

# Recital M2 - HR example

Welcome back for the second example.

Let's now review the HR example that you explored in class with Professor Gladdy.

As a quick reminder, we are now the HR department of a big consulting company and

we're worried about the high number of employees leaving the firm.

Play video starting at ::15 and follow transcript0:15

We want to use HR analytics to understand why employees leave.

And to discover the actions that we can take in order to retain our best

employees.

Let's turn to R.

The first thing you always do is set your working directory.

Once that is done you can clean up the memory of

your current R session by running this line.

We are now ready to load our datasets, and we can do so

using the read.table function that we already know from the SKU example.

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

As always, let's explore our dataset using the str function.

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

What we discover here is that our dataset contains 2,000 observations and

has 6 variables.

What are these variables?

First, the S variable, which is numeric.

As you may remember from the lectures,

the S variable is the satisfaction level on a scale of 0 to 1.

Then we have the LPE variable which is also a numeric, and

which is the last project evaluation by a client on a scale of 0 to 1 again.

Then the NP variable, which is an integer variable and which

represents the number of projects worked on by employee in the last 12 month.

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

Then the ANH variable, which is an integer variable and which represents

the average number of hours worked in the last 12 month by the employee.

The TIC variable, again, which is an integer variable and

which represents the time spent in the company in years by the employee.

The last variable is the Newborn variable.

It will take the value 1 if the employee had a newborn within the last 12 month and

0 otherwise.

Play video starting at :2:9 and follow transcript2:09

Now as we did in the SKU example, and

as we will often do, let's check out some summary statistic for

our variables by using the summary function.

As you can see, the mean value for

the satisfaction level is 0.44 and you can check

out some other statistics that might be able to guide you through your analysis.

Now in order to make your variables comparable to one another,

we need to normalize them.

In order to do that, we will create a copy of our dataset and

we will call it testdata.

Play video starting at :2:47 and follow transcript2:47

And in order to normalize our datasets, we use the scale function.

Play video starting at :2:55 and follow transcript2:55

As was explained in the lectures, the scale function subtracts the mean and

divides by the standard deviation of your variable.

We can now compute the distances of our data points using the dist function.

Play video starting at :3:11 and follow transcript3:11

And as our first argument, we've got our dataset, which is testdata.

As a second argument, we've got method and

we set it equal to "euclidian", don't forget the quotes,

in order to compute the distances using the Euclidian method.

We will store the results in d.

Let's run this line of code.

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

As we did previously,

let's perform hierarchical clustering using the hclust function.

We pass it the distances as a first argument and

set the method equal to "word.D" in quotes as a second argument.

And we store the results in hcword.

Let's decide like you did in class to have four different clusters.

As we did in the SKU example, we can assign our points to our clusters using

the cutree function and creating a new variable

to our original datasets called groups, in which for each observation,

this is going to indicate the group to which the observation belongs.

We can now run this line.

Play video starting at :4:25 and follow transcript4:25

In order to get a nice output table, we need to compute the mean values for

each variable and for each group using the aggregate function.

Play video starting at :4:36 and follow transcript4:36

And passing it as arguments ., which means we want to use our variables,

~ and then groups which is the different groups that we want to have.

The second argument is the data we want to perform this on.

And note here that it's the original dataset, so data=data,

which was the name of our original dataset.

And the FUN argument is the function that we want to use in order to group,

to aggregate the variables.

In here it's the mean function.

We can now run this line.

Play video starting at :5:14 and follow transcript5:14

One thing that would be really nice to have is the proportion of our data

that is in each cluster.

Play video starting at :5:20 and follow transcript5:20

To compute that, we're going to add it to aggdata.

We create a variable called proptemp which computes the number of

observations in each group using the S variable here, but we could have used any.

And the function we aggregate with is length,

which count the number of observations.

Let's run this line.

Let's type proptemp in the command line to see the output.

What we see here is in group 1 there are 793 observations,

626 in group 2, 476 in group 3, and 105 in group 4.

We are now ready to add the proportion variable to our aggdata.

And we can do that by computing the ratio between the number

of observations in each group, which is proptemp$S, and

the total number of observations, which is 2,000 if you remember.

But that you could have the sum, proptemp$S.

Let's run this line.

Now to better organize our output table,

we order the groups from the one with the largest number of observations

to the one with the smallest by using the order variable.

We run the line.

To see our output table, we type aggdata in the command line.

Play video starting at :6:42 and follow transcript6:42

And we can see now our aggdata output table,

with the proportion as the last variable.

Now if you do a view(aggdata),

you can see that one segment has 100% of employees with newborns,

while the other segments have no newborn at all.

This variable is a 0 or 1 outcome and

what we observe is an artifact of a binary outcome in a clustering.

Those types of variable will typically drive the results.

On top of that, as discussed in the videos,

we could wonder if the Newborn variable is really relevant in this context.

So as discussed in the videos, let's remove the Newborn variable.

We create a new dataset called testdata and we set it equal to data.

We include all the rows of the dataset, and

only column one through five, which is all the columns except the Newborn column.

We then rerun the code used above.

If you take some time to read it, you'll see that it's exactly the same thing but

we're calling it on the reduced dataset now.

Assign our observations to the groups.

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

Aggregate the values again.

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

Compute the proportion and order the result again.

Let's see the output by calling aggdata.

Play video starting at :8:3 and follow transcript8:03

At this point,

you should get the same output that what Professor Gladdy showed in class.

A good visualization idea is to save the output in a CSV file to work on it

later in a spreadsheet tool like Microsoft Excel or Google Spreadsheet or the sort.

And you can do that with the write.csv function that you can see right here.

You call the dataset on it.

Your second argument is the name of the file that you want to save.

And our third argument here just indicates that we do not want to have

the display of the row names in the output file.

If for any reason with the write.csv you encounter an issue,

it is most likely due to regional settings with the separator and

I recommend that you use the write.csv2 function instead.

This wraps up our second example, and I will see you in the next video.

# Recital M2 - Telecom example

Welcome back for the third example.

This time, we're a telecom industry company and we would like to find groups

of our clients that use our services in similar ways.

This is essentially an example of customer segmentation.

Let's set our working directory,

clean up the memory of our current R session, and load our data set.

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

As we've done so far.

Let's explore our data set with the str function and

check some simple statistics with the summary function.

Play video starting at ::31 and follow transcript0:31

The output of the str function tells us that we have 1,000 observations and

five variables.

Our variables are the following, first the variable, which is numeric and

which is the average number of hours of polling per month from each

of individual in the data sense.

And then you have the intern variable, which is numeric, and represents

the average number of hours spent by the user in international phone calls.

Then we have the text variable, which is numeric, which represents the average

number of text messages sent by the user every month.

Then the data variable, which is numeric, and

represent the average data usage per month in gigabytes.

The last variable is the eighth variable.

Which is an integral variable,

and which represents the age of the user of the line in years.

Let's take out some summary statistics.

Here you can see that the mean of the text variable is around 225.02.

And the mean of the data variable is about 1.987.

These are clearly on very different scales, and as we've done

before we are going to make our variables comparable by normalizing our data sets.

So we create a copy of our data set called Taze Data, and

then we scale our variable with the scale function.

We now compute the distances, as we've done before with the disk function and

perform your cluster with hclust function.

So let's decide, like you did in class, you have eight clusters so

we create the group's variable that we add to the original data set called Data, and

use the qtree function to assign our points to our eight clusters.

Play video starting at :2:9 and follow transcript2:09

Then in order to get a nice output table, we compute the mean values for

each variable and for each group by using the aggregate

function that we've used before with the Fun argument.

Play video starting at :2:20 and follow transcript2:20

Like we've done in the HR example,

we compute proportions by using the Calls variable here.

And the length function.

Play video starting at :2:29 and follow transcript2:29

And add the proportion by computing the ratio between the number

of observations in each group and the total number of observations.

To better organize our output table, we order the groups from the one

with the largest number of observations to the one with the smallest number.

You can then save your output table by using

the write.csv function that we discovered in the HR example.

For a telecom company point of view,

eight segments might be more than what they actually want.

Let's try again with five segments and

see if some of our segments can be treated similarly.

So back to our code.

We changed the groups variable to qtree hc

word as a first argument and k equal for five segments takes time.

Let's run it.

Let's check out our output table by typing Data.

Now you can save this output with the write.csv function.

Or if you have some issue with the separator,

you can use the read.csv2 function.

So far in this tutorial, we have not used R a lot to create our visualizations.

So, let's see how you can use R to create a rater chart.

To better have utilized the output of our clustering.

This code will allow you to replicate the charts shown by Professor Gladdi

during the lecture.

First, let's select the colors that we want to use with the palette function

to which we pass the rainbow function.

Then in order to create the rater chart, we use the stars function.

Feel free to explore all the arguments of

the stars function by typing $\* in the command line.

Play video starting at :4:1 and follow transcript4:01

In the right bottom corner of my screen, you can now see the output

which is the same rater chart that you obtained in class with Professor Gladdi.

This wraps up our recital for this module.

I hope you learned a lot and will play around with R in order

to solve your own problems and explore your own data sets.

See you next module.

# Script and dataset files to replicate recitals

[R script module 2](https://d396qusza40orc.cloudfront.net/phoenixassets/strategic-analytics-s12n/new/FSBA_SCRIPT_MODULE2.R)

[DATA\_2.01\_SKU.csv](https://d396qusza40orc.cloudfront.net/phoenixassets/strategic-analytics-s12n/new/DATA_2.01_SKU.csv)

[DATA\_2.02\_HR.csv](https://d396qusza40orc.cloudfront.net/phoenixassets/strategic-analytics-s12n/new/DATA_2.02_HR.csv)

[DATA\_2.03\_Telco.csv](https://d396qusza40orc.cloudfront.net/phoenixassets/strategic-analytics-s12n/new/DATA_2.03_Telco.csv)

# Wrap-up: identifying groups within data

In this module, we have explained that the purpose of finding groups, or

patterns, or segments within data in the context of business analytics is to

improve your business efficiency.

That is, to find the right balance between

treating different cases differently to achieve effectiveness and

treating similar cases similarly to achieve economies of scale.

That's why clustering is the perfect tool in this context.

Clustering helps us achieve those objectives by identifying similar cases.

That is minimizing the distance between observations within the same cluster

from dissimilar ones, that is maximizing the distance between clusters.

We have seen three examples.

A simplistic example from supply chain analytics

where common sense could be used to distinguish groups of products

that would be managed similarly in terms of supply chain management.

Second, a multi-dimensional example from HR analytics.

Where we could detect different explanations for

employees who leave the company, allowing us to retain them more efficiently.

The third example is an introduction to customer segmentation,

where we analyzed the consumer's product usage in telecommunications in order

to identify groups of similar habits and serve our customer base more efficiently.

Play video starting at :1:39 and follow transcript1:39

We have seen that what drives our analysis is not technical.

Instead we are driven by the business issues and

we assess whether our results are actionable or not

Play video starting at :1:52 and follow transcript1:52

and whether our recommendations are efficient or not.

Play video starting at :1:56 and follow transcript1:56

For each example, we spent some time to understand how our results could be

reported relevantly for a business audience.

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

When we need to convince others to change the way they do business,

Play video starting at :2:10 and follow transcript2:10

even if the message is complex.

Play video starting at :2:13 and follow transcript2:13

Actually particularly when the message is complex

you need to communicate effectively and clearly.

You can replicate all the examples that are presented in class

thanks to the recital where we provided our scripts and

data to allow you to understand better how it works in practice.

And I would advise that you try to play with the data yourself as much as you can.

# Understanding causes and consequences: introduction

One of the main reason to use business analytics is to understand

the relationship between causes and consequences.

And in the business world where any incremental percentage of

efficiency gains may be a competitive asset.

If you want to be able to act efficiently you need to be able to

anticipate in a quantified way what will be the consequence of your actions.

Play video starting at ::36 and follow transcript0:36

And to obtain this objective,

business analytics is probably the most perfect approach.

In this module, we'll first explain why using rigorous statistical

methods to understand relationships between different events is crucial.

As a manager, you cannot merely rely on correlations.

To anticipate the results of your actions,

you need to use robust statistical methods, instead.

We'll cover two examples.

First, we'll take an example of credit scoring.

How to measure what makes an individual more or

less likely to have a strong credit score.

Then in a second example, we'll take again the example of HR analytics

to understand what makes an employee more or less likely to leave the company.

I could present the most exotic and advanced statistical techniques and

methods to understand causes and effects, but

I will actually do the exact opposite and use standard and classic ones.

Namely, the linear regression and the logistic regression, and this for

two reasons.

First, because like it or not, regressions do actually very well in practice.

The improvement in terms of accuracy of more advanced techniques doesn't really

pay off, because the cost in terms of complexity may be really high

while the accuracy improvement itself is too often only marginal.

Play video starting at :2:6 and follow transcript2:06

And two, at the end of the day our objective is to understand

the relationship between causes and effects, and

to communicate it to an average business audience.

We want to pass from the statistical results to managerial relevant

business oriented insights and the approach I present today actually

applies similarly to very advanced techniques or more basic ones.

Play video starting at :2:34 and follow transcript2:34

So even if you want to use only state of the art algorithms in your own practice,

this module will be useful for you as well since we'll mainly focus on how to

interpret our results with a business mindset and how to report and

communicate our conclusions efficiently regardless of the analytics tools we used.

As we did before, all the techniques and examples presented during the videos will

be applicable thanks to the recital presented at the end of the module.

# Why use Business Analytics to understand the relationship between causes and consequences

As a manager, you want to be able to anticipate what will be the consequence

of your actions, in order to maximize your effectiveness and efficiency.

Play video starting at ::20 and follow transcript0:20

Nowadays the idea that one needs to rely on data to make a decision,

instead of gut feelings or one's opinion is generally accepted.

But too often, people still believe that correlation is the same as causation.

It's not.

It's not because two events are happening at the same time, that you may conclude

that the first one is the cause and the other one is the consequence.

Play video starting at ::45 and follow transcript0:45

As matter of fact, very often both are the consequences of the same cause.

And if you fail to see it, you won't be efficient because you will try to change

something, but will have no consequence on the outcome you want to obtain.

As a first example, let's imagine that we are dealing with credit risk,

and that we're in charge of selecting the customers, that may or

may not be granted a loan.

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

What we are looking for, is actually for

a way to understand how to measure the "credit score" of those applicants.

And then, make a decision based on this scoring.

The first step, is hence to understand

what is driving credit scores in a quantitative and robust way.

Play video starting at :1:32 and follow transcript1:32

If you have hundreds of applicants, we cannot just guess.

It's better to have a rigorous approach that will be statistically efficient.

Play video starting at :1:41 and follow transcript1:41

Let's take an example that is provided in the book,

"An Introduction to Statistical Learning with Applications in R",

from James, Witten, Hastie and Tibshirani.

In this data set, we have 300 individuals who were assigned a credit score.

Play video starting at :1:59 and follow transcript1:59

Actually, in the original data set,

we had some more individuals, but I'll keep those aside for the moment.

And I'll explain why later.

Play video starting at :2:8 and follow transcript2:08

We also have information about the yearly income.

The number of credit cards owned, their age, their level of education,

their gender, whether they are a student or not, whether they are married,

their ethnicity, and the average credit card debt that we'll call balance.

Play video starting at :2:29 and follow transcript2:29

A regression models the relationships between potential causes

(the independent variables), and an outcome, (the dependent variable).

As usual, we leave the technical considerations aside, and

focus on how this tool can help us.

Play video starting at :2:46 and follow transcript2:46

In practice we should focus on 3 aspects.

One, what's the explanatory power of the model?

Or said otherwise, is the model really helpful in describing the situation?

Play video starting at :2:59 and follow transcript2:59

This can be measured with a correlation between the values

as recovered by the model and the actual ones.

Play video starting at :3:7 and follow transcript3:07

Or the number of correctly classified observations

if we use a logistic regression as we'll see later.

If the model explanatory power is weak,

we can not say that the conclusions will be decisive.

Play video starting at :3:18 and follow transcript3:18

But if the model describes the situation well,

we can then focus on two: what are the significant factors?

Play video starting at :3:25 and follow transcript3:25

The first outcome of a regression will be to identify the statistically

significant factors from those that are not.

This will be assessed using the p value of the effects.

Again, in this MOOC I'll not explain what the p value is, but

you should just see it as a measure of how weak the significance is.

So the smaller the p value, the more significant the effect is estimated to be.

Play video starting at :3:53 and follow transcript3:53

We can only actionate an effect that is statistically significant, but

if it is statistically significant,

we can then look at three: the sign of the estimated effect.

Is it positive?

Is it negative?

And what are the most important ones?

Note that - as it is often the case in statistics -

we will first be interested in knowing whether an effect is significant or

not, before wondering whether the effect is positive or negative.

Play video starting at :4:21 and follow transcript4:21

Because at the end of the day,

significance without knowing whether the effect is positive or

negative, means that we should take care of the driver as a manager.

Play video starting at :4:32 and follow transcript4:32

While only knowing whether the effect is positive or negative, without knowing

whether the effect is really significant, is not actually actionable in practice.

So it's indeed significance that is of primary interest!

All this information can be provided with our R tools,

such as the lm function for instance.

So let's apply it to credit scoring situation.

The first thing we need to assess is whether the model does a good job or

not at estimating the credit scores.

As we'll see during the module dealing with predictions, the most rigorous way to

assess the quality of our estimator, would be to take out of sample data and

compare estimations with actual results for

observations that have not been used to estimate the model parameters.

This would assess whether our estimator does a good job,

at "predicting" the outcome in an unknown situation.

But here I'll just look at the "in-sample" accuracy.

For each observation the regression provides a credit score that is the one

"estimated by the model."

Play video starting at :5:38 and follow transcript5:38

Those are what we call the fitted values.

We can then compute the correlation between those "fitted values" and

the actual credit score.

Play video starting at :5:49 and follow transcript5:49

You will see that you obtain a correlation of 99% which is amazingly good!

But we need to be careful with correlations,

because it may be driven by the most extreme observations.

Imagine that you would have only one accurate prediction.

But that it would be a very extreme one.

Say, 1,000 times bigger than the other ones.

In that case, we would have a correlation close to one as well.

So let's find a visual way to assess the quality of predictions

by plotting the fitted values as a function of the actual credit score.

Play video starting at :6:22 and follow transcript6:22

And we see then that the recovery isn't very good.

Play video starting at :6:26 and follow transcript6:26

Even if, as I just mentioned, the smaller values are not as well estimated

as one could believe with a correlation of 99%.

Now that we are confident in the quality of our model,

let's look at the significance of the effects.

The summary function provides all the relevant information about the model.

In particular, it reports the level of statistical significance,

with three stars when it's very strongly significant, two when it's significant,

and only one when it's weakly significant.

A dot means that the significance level is very low.

Play video starting at :7:2 and follow transcript7:02

No symbol would then mean that it's not significant at all.

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

The intercept is not really relevant here.

And we see that we have income, balance, and

the fact that the applicant is a student or not that are very significant.

Play video starting at :7:18 and follow transcript7:18

The age is only weakly significant.

Play video starting at :7:21 and follow transcript7:21

Now let's focus on the significant drivers and look at their effects.

What are the most important ones?, and is the effect positive or negative?

Play video starting at :7:32 and follow transcript7:32

The problem is that the drivers are not directly comparable,

some variables are zero or one, some are categories and

some are continuous variables with different ranges.

Play video starting at :7:43 and follow transcript7:43

So, let's look at the absolute value of the T-value column.

This column allows to assess the strength of the significance of a driver.

Play video starting at :7:54 and follow transcript7:54

The larger the absolute value, the better.

Play video starting at :7:56 and follow transcript7:56

We see that the most important driver is balance,

then income, then the fact that the applicant is a student.

And only then, the age which is only weakly significant.

We can then look at the sign of the estimates, which

indicates whether the effect is positive or negative on the credit scoring.

Play video starting at :8:17 and follow transcript8:17

Finally let's only report the relevant information for a manager, and

let's do it in a visual way by just focusing on how important the factors are.

By ranking them and categorizing them with colors, whether they are positive or

negative

we then get this table.

Play video starting at :8:36 and follow transcript8:36

A table is nice, but

can't we think of a way which is even more visual to report those effects?

Play video starting at :8:42 and follow transcript8:42

We can use a plot maybe.

For instance,

we could report the dependent variable, here the credit score, as a function

of the most important factors we've identified thanks to regression.

If you do it in R, you will see that you can then obtain this plot.

Which reports nicely the relationship between the dependent

variable, and the balance.

Which is, as we have seen, the driver impacting the most credit score.

As long as it's consistent with the effects estimated in the regression,

you can always produce as many plots as needed to report in a visual way

how the causes impact the consequence you are interested in.

People will always will understand easier,

a plot like this than a story about p-values and t-tests.

For instance, you could also do the same thing with the income, like this.

Play video starting at :9:33 and follow transcript9:33

But never forget, that it's only an "unconditional" result, meaning that

you do not take into account the other effects here, where the regression does.

Play video starting at :9:44 and follow transcript9:44

But as long as it's consistent with rigorous statistical approaches,

this type of representation should always be preferred in practice

since it helps your audience to understand the message you want to pass.

# Understanding what distinguishes two categories

Let's now discuss the situation where we need to

understand what makes the difference between two categories.

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

And let's take again the HR analytics example we've seen previously.

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

There we can investigate the question of what's driving the employees' attrition.

First, you know that to understand what's driving this attrition, we need to add

dataset used in the last module, the employees who didn't leave the company!

Play video starting at ::39 and follow transcript0:39

We cannot understand what's making a difference between two categories

if we observe only one of them.

So let's add 10,000 employees who didn't leave to the sample.

Play video starting at ::53 and follow transcript0:53

We use the same variables as before: namely, satisfaction of the employee,

her project evaluation, number of projects done, utilization,

time with the company and whether the employee had a baby recently or not.

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

We can compute that we have around 17% of employees who left this year.

This is huge!

As I'll explain in a future module, this type of information should be at

the very beginning of a presentation on HR analytics for this company.

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

This is a huge issue to solve, and reporting this value would raise

the awareness for the relevance of conducting an HR analytics investigation.

Play video starting at :1:36 and follow transcript1:36

Then, one could first think to perform a correlation with the cor command in R again

and see what are variables that are correlated with the fact that an employee

left the company.

During the recital, you will see that we can then obtain these correlations.

We see that the satisfaction is

negatively correlated with the departure of the employees.

The evaluation is weakly positive and so on.

Play video starting at :2:6 and follow transcript2:06

Here we have each variable considered separately.

Whereas it may be that "given a certain level of satisfaction, the number of projects

done" would have a different effect than the positive correlation we see here.

Play video starting at :2:20 and follow transcript2:20

Actually, we see later that, in this case, everything else being equal,

the number of projects done DOES have the opposite sign.

It has a significantly positive effect on retention,

hence a negative relation to attrition.

So, to address this issue rigorously we can use a logistic regression, for instance.

We'll do it in R by using the glm function.

A logistic regression produces a similar output as a linear regression,

except that we're interested in an outcome that is 0 or 1 and not continuous.

Once we've estimated the relationship between the event

"left" which indicates whether a employee left, (that is a 1), or

stayed (that is a 0)

with the other variables, we can first assess the accuracy of this model.

We can compute the proportion of correctly classified observations for instance, and

you will see that we managed to correctly classify 95% of the loyal employees.

But only 19% of those who left.

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

There are few employees who left,

so it's more difficult to detect.

As we'll discuss during the recital, you can decide to decrease the threshold

of the probability as from which an employee is considered as likely to leave.

That's actually called the cutoff.

But let's focus on the interpretations for the moment.

Overall, we currently classified 82% of the observations.

This can be seen as acceptable and

we can consider the model reliable.

In practice, deciding of what is an "acceptable" classification accuracy

is very often relative.

If you have no clue,

even something a bit better than a random would already be an improvement.

Play video starting at :4:3 and follow transcript4:03

Let's look at what are the significant values

using the last column of the summary command.

Play video starting at :4:10 and follow transcript4:10

Remember that it reports the level of statistical significance

with three stars when it's very strongly significant.

Two, when it's significant and only one when it's weakly significant.

No star would mean it's not significant at all.

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

Here, we see that everything is significant!

Play video starting at :4:29 and follow transcript4:29

This is probably due to the fact that with 12,000 employees in the dataset,

we have a lot of observations.

And something that statisticians know very well is that,

as the number of observations increases, the effects all tend to be significant.

As a matter of fact, when you're dealing with very "big data", say millions

of observations, statistical significance usually becomes meaningless.

Because even for effects that have in practice no impact, a regression

will probably find it significant statistically.

Anyway! Since all of the effects are significant, let's focus instead on:

One, how important they are.

And two, if the effect is positive or negative.

Play video starting at :5:18 and follow transcript5:18

We can use absolute value of the z-value column to assess the importance

of the variables.

As with the t-value for the linear regression

in the case of a logistic regression, the larger the z-value,

the more important the effect is.

And we see that the most important variable is the Satisfaction first.

The time with the company second.

Then the number of projects, and so on.

Play video starting at :5:44 and follow transcript5:44

Let's hide those effects and report them in a self-explanatory way.

As for instance with this table.

Again, I decided to report the most important effects first and use colors.

Where green is for positive effect, since green is usually seen as positive, and

red as negative.

Play video starting at :6:3 and follow transcript6:03

You probably noticed that, when I say positive, I mean it in the business sense.

It has a positive impact on the retention of the employee.

And that's what you want.

If you look at the estimates, it is as a matter of fact, negative.

But since the statistical estimate itself is probably meaningless for

most of your audience - because usually, like it or not,

most of the business practitioners are not that interested in statistics methodology -

we replace a statistical indicator by something that we'll convey our

message better.

So that's why here we use colors that are generally considered as positive for

positive impact on your business.

Let's wrap-up what we've done until now.

What did we do to report the results of the regression?

One, we provided some numbers reporting clearly the accuracy of the model.

Play video starting at :6:59 and follow transcript6:59

Here we distinguished between two types of errors:

when our model estimated that someone should have left but she didn't

(that's a false positive.)

and failing to estimate correctly that someone has left

(that's a false negative). But we didn't use statistical terms and

focused on the business interpretations instead.

Play video starting at :7:23 and follow transcript7:23

We also provided an overall head of correctly classified observations.

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

In a second step,

we reported a table focusing the effects that are significant or not.

In red, we report the "bad effects."

Those impacting negatively what we want the outcome to be:

here we want to retain the employees.

And in green, we have the "good effects."

We assess "good or bad" in the business sense, not the statistical sense.

Play video starting at :7:54 and follow transcript7:54

Now, in reality,

we should be really careful when interpreting the result of a regression.

Because very often, the relationship between the outcome variable and

the explanatory variable, is not as unidirectional as it may seem.

Play video starting at :8:9 and follow transcript8:09

Satisfaction certainly explains the fact that you want to stay or

not with the company.

But the relationship between your evaluations and staying at the company, for

instance, may go two ways.

The company may decide to fire an employee if she has poor evaluations.

Play video starting at :8:26 and follow transcript8:26

But if another employee decided to leave already,

and that she's looking for the job,

it may impact her performance.

And there may be a delayed effect where the cause is the anticipated departure,

leading to a decline in motivation,

and the final consequence is a poor evaluation.

Play video starting at :8:48 and follow transcript8:48

This "loop of causality" will cause what we call "endogeneity" in statistics.

And while,

I won't explain what it is, because it's clearly out of scope of this training,

I can tell you that it will result in your effects being estimated inaccurately.

So be careful when interpreting the effects you estimate.

And see your results as clues, leading you to your destination, but

not as decisive facts.

So let's investigate in this further

how the effects we identified are really

related to the employees leaving the company in the next video.

(Required)

English

​

Help Us Translate

# Beyond the regression estimates: reporting effects in a visual way

In the previous video, we analyzed an HR analytics example and

we showed that the two most important effects explaining that the employees were

leaving the company

were the satisfaction and

the number of years since the arrival of the employee in the company.

We used a table to report those effects but

you can do better, we can use a plot, instead, to report our effects.

And actually,

now that we know what are the important effects, we can even use a method,

that does not rely on statistics to help the audience to understand the situation.

This is called a Model-free approach.

Let's first state the example of the "Time" an employee spent within the company and

it's relation to attrition.

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

Since we're willing to find a visual representation of the relationship between

these two elements

we could just use the plot function in R and

report the fact that the employees left or not

as a function of the time spent in the company.

But if you do that without preparation we obtain this plot and

we directly see a problem.

Since the variable is binary it's either 0 or 1 you have at least one, 1 and

at least one0 for every year we considering from 2 to 6.

The problem then is that we don't know how many observations are zeros and

how many observations are ones.

So instead, we could aggregate the values and

compute the mean attrition rate by years spent with the company.

We then obtain this graph, which is a lot better.

What do we see?

As time goes by employees are - as expected - more likely to leave the company.

The attrition rate increases from year to year.

In year 2 the attrition rate is close to 0 but in year 3 and

4 it goes respectively to 17 and then 24%.

And after 5 years employees are on average around 44% likely to churn!

And it's not correlative.

It basically means that if you pick 2 employees who are at the end of their

4th year randomly by the end of their 5th year one of them probably left.

Play video starting at :2:21 and follow transcript2:21

We should definitely investigate why we cannot retain those employees.

Play video starting at :2:26 and follow transcript2:26

Is it because the salary doesn't increase steep enough?

Is it because we fail to propose a decent career plan,

Play video starting at :2:34 and follow transcript2:34

it may exogenous, it may be somehow out of our hands.

For instance because employees didn't

intend to stay more than four years in consulting anyway.

Play video starting at :2:45 and follow transcript2:45

But in any case it should be investigated.

And we should do our best to at least retain the employees that

are performing well.

We'll discuss how we could prioritize our actions during the next module.

There's something strange though.

We see that in year 6, while the retention rate is still dangerously high

it decreases a bit compared to year 5.

Play video starting at :3:7 and follow transcript3:07

What does that mean?

Could it be that there is a non linear and non monotonous effect?

Meaning that it starts increasing and then it decreases.

Or could it be that it's just noise in the data?

To assess it we can compute the number of observations related to each dot.

Play video starting at :3:25 and follow transcript3:25

And we could for instance make the same plot as before but

with the size of the bubble proportionate to the number of employees in that case.

We then obtain this plot and

we see that the bigger group is actually the three year old employees and

that the small group is indeed the 6 year old employees.

On the other hand, if you compute the exact number you can see that

this group has 512 observations.

Even though it is the smaller one,

we cannot really say that it should be disregarded.

It may actually be an interesting case of self selection, those who stay

even after passing threshold of 5 years are those who will stay even longer.

Play video starting at :4:13 and follow transcript4:13

It's something that we very often see in marketing as well.

At some point your customers that are still loyal to you after many years

will probably stay loyal forever.

The same with employees...

Play video starting at :4:25 and follow transcript4:25

Now, let's focus on another factor.

The one we identified as the most important satisfaction.

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

This case is a bit more complex because in contrast with the time spent,

that took only five different values, satisfaction, as an index and

can take any value from zero to one.

So, to be able to plot it on a graph in the same fashion as what we did before,

with the time spent, let's rank the satisfaction from the most satisfied

employees to the least satisfied ones.

And let's compute again the average attrition rate

by groups of similar satisfaction ranking.

Play video starting at :5:2 and follow transcript5:02

We then obtain this plot.

Play video starting at :5:4 and follow transcript5:04

Let's look into the details because here we observe

very interesting non-monotonous effects.

Play video starting at :5:11 and follow transcript5:11

First, we have the very satisfied employees.

They report a low risk of leaving the company and

we also have what I'd call the "it's okay" employees.

Play video starting at :5:22 and follow transcript5:22

Not very satisfied but they're okay and they won't move.

Well in-between you have something very strange.

Play video starting at :5:30 and follow transcript5:30

The employees that are satisfied but still not enough

seem to present a higher risk of leaving than other employees that are more or

less satisfied.

It's probably those employees we've discussed during the second module on

"finding goals within data".

Good performers, but either they worked too hard and were burned out or

they were probably recruited directly by the client.

In the case of the burned out, they were not as happy as they could have been so

they left.

Play video starting at :6:4 and follow transcript6:04

And in the case of the high performers,

their level of satisfaction was not high enough to retain them from leaving.

In any case those are probably the guys we should do something about.

Play video starting at :6:17 and follow transcript6:17

As a follow up exercise, I would advice you to look at what's their level

of performance using the evaluation table.

Finally we have the very unsatisfied employees. As expected, those

guys report a very high level of attrition and we should do something about it.

Play video starting at :6:35 and follow transcript6:35

Understanding why they are unsatisfied if solvable

may reduce the employee attrition of the company.

But we should first check that we are sure we want to keep him,

again the evaluation variable should be included in the reflection.

Play video starting at :6:52 and follow transcript6:52

But I'll let you play with the data yourself.

Play video starting at :6:56 and follow transcript6:56

Actually, since we observe highly non-linear effects a

more advanced models, such as decision tree or ensemble methods,

would probably perform a lot better than a simple logistic regression.

I invite those of you who are eager to use those techniques

to practice on this data set.

But don't forget to avoid the black box syndrome: a very accurate model is nice

but when you're trying to understand a situation, an insightful model is better.

Play video starting at :7:27 and follow transcript7:27

And the challenge with more advance models

is usually that their output is more complex to interpret.

To wrap up on the HR Analytics case

I'd like to emphasize how we came to produce the results we presented.

We first build a model and make sure that the model is reliable.

Play video starting at :7:45 and follow transcript7:45

We then identify the most important factors and

report the relationship with the outcome visually.

Play video starting at :7:54 and follow transcript7:54

We finally use a model-free representation of our results.

You may have noticed that in this video

we didn't directly use the results of the regression for the graphs.

The regression was used to identify the most important factors and

only then, we reported visually the effects of those factors but

without using the estimates directly anymore.

Play video starting at :8:20 and follow transcript8:20

Since, what we saw was consistent with the results of the regression, it was reliable.

But thanks to this model-free approach, we could go further in the interpretation -

and hence or conclusions -

using visual clues which allows to observe complex effects.

And it didn't require that your audience understands anything about statistics.

Play video starting at :8:44 and follow transcript8:44

This approach could be replicated to any type of sector, any type of application,

and any type of variable.

Also with more advanced models is the combination of statistical techniques

with the articulation of insights presented visually that will allow

us to build a shared representation of the reality with your audience.

As we will discuss at the end of this training this is key to ensure the success

of any change you want to implement for your business.

# Recital M3 - Credit score example

Welcome to the recital for module three.

This week, we focus on understanding the relationship between causes and

consequences, in order to be able to quantify the consequence of our possible

future actions.

Our main focus will be to try to understand like you did in class,

what drives our credit score.

Let's get started in r.

You're now familiar with the r studio.

And the first thing you do is to set your directory

to the folder where you have downloaded the credit scoring data set.

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

To make sure that the memory of our current r session is clean,

we run the fulling line.

We then load our data set.

Play video starting at ::40 and follow transcript0:40

By now, you should be familiar with the arguments, but

if you do not remember what they're used for,

remember that you can always type the name of a function in the command line,

preceded by a question mark in order to have access to it's documentation,

which includes the description of its arguments.

Here we would type ?read.table and hit Enter.

Play video starting at :1:5 and follow transcript1:05

Let's now have a look at our variables with the STR function and

see some summary statistics with the summary function.

Play video starting at :1:15 and follow transcript1:15

The STR function outputs a description of our variables.

We find out that we have 300 observations and 10 variables in this daily set.

First, we have the income variable

which represent the income in thousands of US dollars.

We then have the rating variable, which represent a credit score.

Then the cause variable, which is the number of credit cards that the user owns.

Then the age variable representing the age of the individual.

The education variable shows the number of years of education of the individual.

The gender is a factor that says male.

If the individual is a male or a female if the individual is a female.

The student variable is again a factor and tells us whether or

not the individual is a student.

The married variable tells us whether or not the individual is married.

The ethnicity variable provides us with the ethnicity of the individual,

for example, African American, Asian, Caucasian, etc.

And the last variable is the balance which corresponds to the average credit card

debt for the individual.

To have a better understanding of our data and have in mind the order of

magnitude of our variables, we run this summary function on our dataset.

We can see, for instance, that the minimum income value is $10,350.

The maximum is $186.63 thousand,

while the mean value is $44.05 thousand.

Similarly, the rating variable has a minimum value of 93,

a maximum value of 949 and a mean value of 348.1.

This gives us a good idea of the order or managing of the rating variable.

In this data set, our dependent variable is the rating variable.

Or in other words,

it is the variable we want to predict based on our other variables.

Before we move onto any kind of prediction making and modeling,

we would like to know how our rating variable is distributed in our data set.

A histogram is a great graphical representation of the distribution

of a numerical variable.

To build a histogram in r, you use the hist function and

you pass on as an argument, the numerical variable that you want to explore.

If we run this line,

we get the histogram that you can see on the bottom right corner of my screen.

Let's explain the output quickly.

We see the frequency on the vertical axis and

the rating variable on the horizontal axis.

What the code has done is to define buckets,

which contain credit scores between zero and a hundred.

Between 100 and 200, between 200 and 300,

etc until the last bucket which contains credit scores between 900 and 1000.

What do we see visually?

Well, we see that the distribution is skewed to the right.

That means that while most our values seem to be between 100 and about 400.

We have created scorings way beyond up to our last buckets.

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

Now, what we're trying to do is to understand the main drivers of

the credit score.

One thing we can do, is to have a look at the correlations between our variables.

As explained in class, it only makes sense to do so with numerical variables.

To obtained the correlations over variable in art, we used a core function,

that you can see here and pass it on as argument or data sets.

So data, excluding the catigurable variables,

Play video starting at :5:3 and follow transcript5:03

which are if you remember gender, student, married, and ethnicity.

If you look at our data set this contigurable variables are in

the sixth to the ninth column, so we only set it column one to five and

column ten of our data set by using the combined or C function.

As explained in class, the output of the correlation function provides us

with numbers that do not really tell us which variables are most important, or

significant in predicting the credit score.

We do not know either how they contribute to that score.

Play video starting at :5:38 and follow transcript5:38

Is it positively or negatively?

So as you did it in class, we'll use a regression and

learn how to create a regression model in r.

To create a linear regression model in r, we use the LN function.

Play video starting at :5:52 and follow transcript5:52

Our first argument is the values we input to build a model.

So the rating variable comes first as it is the variable we are trying to predict.

Then, you insert a tilde and

to tell r that we want to use all the other variables we insert a dot.

Note that if you wanted to use only some of the variables to build your muddle, you

would write the name of those variables and separate each of them by a plus sign.

Our second argument is to say data

equal to the data sets that contain our variables.

Here, our dataset is called data so we said data equal to data.

Now that we have a model we should find out how it performs.

As in class, we'll stick with in simple accuracy.

Note that a more rigorous approach would be to test the model and

data it has never seen before.

Just to be clear, in simple data is the data we use to build the model.

In our case it is our data daily set.

For each of our observations, the model estimate the value of the credit score

based on the other variables and these are called the fitted values.

To assess the model on in sample data, you can use correlation between the actual

values and the fitted values, which again obtain by going the correlation function.

As you can see, we obtain 0.90, almost 0.99 as correlation coefficient.

But as explained in class, correlation has some limitation.

To circumvent these limits, we can use a more entrative and visual approach and

plug the fitted values as a function of the actual values.

Let's run the line.

We can clearly see that the model does pretty well.

That said, it does not do as good of a job on the lower values.

In fact, it seems to be pretty off with the lower values.

Once we're confident in our model,

the next thing we can do is to check out the output of our regulation model.

To do so in r, you use the summary function on your model like this.

You learned in class how to quickly integrate the results of a linear

regression, but let's go very quickly over some of our main take aways here.

First, the star system in r tells us that the statistically significant variables

are the income variable, the student variable, and the balance variable because

the absolute value of the T value of the bounds variable is the largest.

We conclude that it's the strongest driver.

Then the income, and then whether or not the applicant is a student.

Finally, we want to know whether the impact on the rating is positive or

negative.

For each.

We can find that information in the estimate column and

surprisingly, a negative coefficient indicates a negative impact

while a positive coefficient indicates a positive impact.

So here, the fact that the applicant is a student, impacts negatively on the score.

In other words.

It lowers the score.

On the other hand, the positive coefficient next to the balance indicates

that the higher the balance, the higher the credit score.

Same for the income which is pretty intrusive.

Now, if we wanted to show our technical fundings to business students,

we could plot the rating as a function of the balance.

And show that the slope is positive and

that indeed, the higher the balance the higher the rating.

If we quickly run the same line with the income instead of balance

we can also see visually that the higher the income, the higher the rating.

That's it for this example.

I hope this made things clearer for you and I will see you in the next video.

# Recital M3 - HR example

Welcome back.

Let's now turn to the second example explored in class this week,

namely the HR example.

As a quick reminder, our goal is to understand what distinguishes

employees who stay in the company, and those who leave the company.

To do so we're going to explore a data set very similar to the one we

had in module two.

However, this time, we've added more employees, and more importantly, we've

added an outcome variable which tells us if the employee indeed left or not.

All right, let's turn to and analyze our data set.

Once you've set your directory to the folder where you have downloaded the HR

data set, you run this line to clean up the current memory of your R station.

We then load our data set with the read.table function and call it data.

Now let's have a look at our variables with the str function and

see some similar statistics with the summary function.

The str function shows us that we have 12,000 observations in this data set,

and seven variables.

First, we have the satisfaction, then LPE,

which stands for last project evaluation, then NP,

which is the number of projects worked on by the employee in the last 12 months.

Then the ANH, which is the average numbers of hours worked

weekly by the employee over the last 12 months.

Then the TIC variable which is the time spent in the company.

And then the newborn variable which tells us whether or

not the employee had a newborn within the last 12 month, and

finally, our outcome variable which is called left.

Which will take the value one if the employee indeed left and zero otherwise.

To have a better understanding of our data sets, we run the summary function.

Play video starting at :1:45 and follow transcript1:45

We see that nearly 17% of our employees have left the company.

We can also see that 15.4% of the employees

had a newborn within the last 12 month and also that, on average,

employees have been in the company for 3.2 years.

Remember that using the summary function and

looking at its output allows you to get more familiar with you data set.

You can often find out a lot of information with the summary function,

which will save you a lot of time and drive your research.

Another thing we could do in order to get familiar with our data sets is to use

the table function in order to obtain the frequencies of the left variable.

Here we see that 10,000 observations are employees who stayed in the company.

While 2000 of them are employees who left.

If we wanted to have the same information but in percentages, we would need to

divided by the end row function on our data set, which counts the number of rows.

Let's run the line, and now we can see that we obtained the same

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

point 16.666% that we obtained with the summary function.

Alternatively, we could also plot a histogram with the hist function

on the left variable and we would see the same information, but visually.

Now that we're more comfortable with our data set, let's turn back to our focus for

this tutorial, which is to try to understand

what is different between employees who leave the company and those who stay.

We can check out the correlation between our variables.

As we said in the previous video,

we call the core function on our data set and run it.

As a quick reminder, correlation only gives us information about the strength

and the direction of the linear relationship between two variables.

Play video starting at :3:33 and follow transcript3:33

If we look at the correlation of all our variables with their left variable.

We see that satisfaction is negatively correlated with it.

TIC is very moderately positively correlated with it.

While a newborn is fairly weakly negatively correlated with it.

Now, this does not tell us much, because we're looking at the relationship

between the left variable and all the others, but separately.

What we want is to understand how they interact with each other.

For instance, here, the correlation between the employee leaving the firm and

the numbers of projects worked on is very,

very weakly positive looking at correlation only.

However you saw in class that all else equal the number of projects

worked on actually has a significant negative relation to attrition.

Looking only at correlation, this is something we would have missed.

Which is why we need to add another tool to our current toolkit and

learn how to build a logistic regression model in R.

To build a logistic regression model in R, we use the glm function.

Our first argument is the outcome variable, for us it's left.

You then insert a tilde followed by the variables you want

to use to build your model.

In our case, we'll use all the other variables, so we type a dot.

We then set the family argument equal to binomial (logit) and

then the data argument equal to our dataset, which is data.

Let's run the line.

First, let's note that our fitted values are the output of the model

on the in sample data.

In the case of a logistic regression model, the output is a probability.

Let's see the proportion of employees' attrition according to the model

by building a histogram of the fitted values.

To do so we use the hist function that we already know and the fitted values.

The frequency is on the vertical axis while the fitted values,

which are probabilities, are on the horizontal axis.

Now let's assess how the model is performing.

We can start by assessing the correlation between the estimated attrition by

the model and the actual one by 0.41,

that's a linear relationship between the two that is positively shown.

That's a pretty good sign.

Now, since the output of a logistic regression is a probability,

that is a continuous variable that can take any value between zero and one.

We need to set a value that tells us what we consider to be a leaver and

what we consider a person who stays.

This is called the cut off value.

Let's set our cut off to 0.5.

This means that any prediction that is above 0.5 will be a leaver,

and any prediction below will be a stay.

Let's evaluate the predictions based on our cutoff value.

To obtain predictions we could run these three lines and we will in a minute.

But for now, let's try to build the confusion matrix in the command line.

What we want is to know when the fitted value is above 0.5.

This will return true.

And if it's below, it will return false.

And then we wan to compare to the actual values.

So we type table logreg$fittedvalues superior or

equal to cut off, and

then we compare with the actual values.

Data.$left.

Let's run it. Now what we see here is the fitted values

and here the actual values.

So here, this means that these people were predicted to stay and, indeed, stayed.

While 536 people were predicted by the model to leave and, in fact, they stayed.

Now based on our matrix,

let's compute the percentage of correctly classified employees who stayed.

Stayed means that the left variable is zero and that the model predicted a false.

Play video starting at :7:43 and follow transcript7:43

We get 0.9464.

Now if we ran this line, we get the same thing.

If you want to replicate this code, I encourage you to use this line

instead of the matrix, which I use only for pedagogical reasons.

Now if we compute the percentage of currently classified employees who

actually left, we're looking here.

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

And if we run this line again, we get the same thing.

Let's compute the overall percentage of correctly classified employees.

So the correctly classified employees are these ones.

And those ones.

Play video starting at :8:27 and follow transcript8:27

Over the total.

Play video starting at :8:34 and follow transcript8:34

And we get 82% accuracy.

Again if we run this line, we get the same thing.

One question you may have is how do we chose a cutoff value?

Well, the answer is not easy.

It depends on the circumstances and the objective of your analysis.

Let me invite you to play with the cutoff value,

in order to understand how it works.

For now, let's set it at .7.

Let's compute the percentage of currently classified employees who stayed.

And here you can see that we get a number that is higher, 0.9805.

Well, it was 0.9464 with a lower cutoff value.

Let's compute the percentage of correctly classified employees who left.

Play video starting at :9:17 and follow transcript9:17

Now we get a much lower value.

We get 0.012.

Well, we had 0.1905 before.

Now let's compute the overall percentage of correctly classified employees,

and we get a lower value at 0.8190 etc., while we had 0.8204 before.

What we find out here is that when the cutoff value is large,

we rarely predict the outcome.

In our case, we rarely predict that the employee is leaving.

But this allows us to identify the employees most likely to leave.

And we can take actions designed specifically to these people.

Now, let's set our cutoff value to 0.3.

First, let's compute the number of correctly classified employees who've

stayed.

Now we get 0.8888, which is slightly lower than

the 0.9464 that we got with the 0.5 as a cutoff value.

We then compute the percentage of correctly classified employees who left,

and we get .473, which is higher than the 0.8,

which is higher than the .1905 that we got before.

What we understand here is that,

if the cutoff value is low, we rarely predict that the employee is staying.

And we make more errors where we predict that the employee will leave.

When in fact, he or she does not leave.

This allows us to identify the employees who might leave.

Here it's really good because we can take preventative actions to try to

retain those people.

Now, that we know our model is performing, let's use the output of

the summary function to understand what distinguishes our two categories.

To do so, we call the summary function on our model.

If we look at the p value, all our predictors are statistically significant.

To look at how important they are, we can look at the absolute value of the z value.

It tells us that the satisfaction level is the most important,

followed by TIC and the numbers of projects worked on.

We then look at the coefficient to see the direction of the effect.

The effect of satisfaction is negative on attrition, so

it is positive for the business.

Same for the number of projects worked on.

This means that the more projects worked on, the smallest the attrition.

Inversely, the effect of the time spent in the company is positive in attrition,

so it is negative for the business.

Now that we know that the time spent in the company, or TIC,

is significant, we want to further explore it.

Maybe we can try to plot attrition as a function of time spent in the company.

To do so, I use the plot function and my first argument is the TIC variable,

which I want to have on the horizontal axis.

While my second argument is our outcome variable, which is our left variable, and

which I want on the vertical axis.

I then added a title with the main argument and

axis labels with the Y lab and the X lab arguments.

Let's now run this line.

Play video starting at :12:20 and follow transcript12:20

Huh.

The output is not really what we intended to have, right?

Well, the reason is that our left variable is binary,

meaning that it only takes two values, namely 0 or 1.

Can you think of a way to circumvent that issue?

Well, what I thought about is to compute the proportion of leavers

by years spent in the company.

That means that that we can calculate the proportion of attrition among all

those who have spent one year in the company, then repeat for

those who have spent two years in the company, etc., etc., up to six years.

How do we do that in R?

First, let's make a copy of our data set and store it in ten data.

Play video starting at :12:57 and follow transcript12:57

Since the left variable is binary, we can use the aggregate function and

compute the main attrition for each value of the TIC.

To do that we use the aggregate function.

Our first argument is the left variable, which is the one we want to aggregate.

Then tilde, then TIC, which is the variable we want to aggregate

on the basis of and then our data set data equal ten data and the function we

want to use as an aggregate here we want to do an average to mean.

Let's look have a look at the output of ag b time rank.

Play video starting at :13:32 and follow transcript13:32

The output is composed of two column.

First the TIC variable and second the mean of the left variable.

The TIC variable takes a value 2 and for the people who have spent

two years in the company, the mean attrition level is 0.01.

Then 3 for the people who have spent three years in the company.

The mean attrition is 0.16, etc., etc.

Up to people who have spent six years in the company,

which have a mean attrition of 0.21.

Let's plot the left column of ag b time rank as a function of

TIC of ag b time rank.

I've added title in labs that you can explore on your own.

So here we find out that attrition increases with the years spent

in the company up to year five, then, it decreases sharply.

This is an interesting fact, so let's investigate further.

We computed mean values, but

we do not know how many employees each dot represents.

To compute the number of leavers in each group, we rerun the aggregate function.

But we change the FUN argument from mean to length,

since we want to do some counting.

And we store the result in CNTB time rank.

Let's check out the output of CNT time rank.

There are 3,021 employees who've spent two years at the company.

5,322 who spent three years at the company, etc., etc., up to our point

of interest, which is the employees who spent six years in the company.,

and there are 512 of them, so there's still a lot of people.

And it's worth reporting to decision makers.

If we want to build a report on it,

we can build a nice visualization by running the following line.

Play video starting at :15:22 and follow transcript15:22

What we're doing here is that we build the same plot as we did before.

What we're doing here is that we build the same plot as we did before and

then we ask r to represent the dots by circles of sizes

proportional to the number of employees in each group.

I won't go into more detail, but

there are a few arguments in this line that you should explore and play with.

Let's now shift our focus to our most important driver of attrition,

namely the satisfaction level and

see if we can get some insights that will allow us to retain more employees.

First, as we did for the TIC variable, we need to prepare our data and

find a way to have a limited number of values that can be taken by satisfaction.

What we can do is to rank the satisfaction variable from the largest to the smallest.

Meaning from the most satisfied employee to the least satisfied one.

As always, first let's make a copy of our data set.

We then add a variable to our team data set called rank status.

We use the rank function because we are ranking from the most satisfied to

the least satisfied, we add a minus sign before the variable we want to go rank on.

Maybe here 10 data, they'll assign S.

We then divide the output by 600 and round the results.

They have 21 groups of similar satisfaction.

Let's run this line.

Let's check out the output by typing teamdata$rankstatus.

What we see is that we obtain, instead of the satisfaction level that we had before,

numbers between 0 and 20, which allows us to have 21 groups of similar satisfaction.

Like we did before, we compute the average attrition rate for each group and

we count the number of employees in each group.

Then like we did before, we plot the attrition level.

Let's see what we got.

Play video starting at :17:16 and follow transcript17:16

Here we have the very happy people that want to stay.

Here we have the people that Professor Grady called the it's okay people.

They're not really happy but they don't really want to leave.

What's way more interesting are those in between.

They're happy but they still want to leave.

This is definitely something you should investigate further if you

were on an HR analytics team.

Maybe they were just burned out.

Maybe they were hired by clients.

In any case, these are likely to be the people you want to target

when undertaking retaining actions.

And then to the right you've got the unhappy people that indeed want to leave,

which is much less surprising.

So that's it for this recital.

I hope you learned a lot and had some fun playing with the data.

I will see you in the next module.

# Script and dataset files to replicate recitals

[R script Module 3](https://d3c33hcgiwev3.cloudfront.net/_0d2c90b40b32694737fc879a92bef9b8_R-script-Module-3.R?Expires=1703721600&Signature=LPxBTGHRuP~n5csmIjzwb-C~pGFsqqqMQwpv8jh8~BCrlx4zySXqzItPTepTuoVyzbeHmckidHs95XFUfkEWnfVds3hGy104ux~fl4x9BDfODWVMlRAPQwKwaTApK29JJPu7TPagiWZHniKnYKzoatpFyNsg-NEoWa74i9j5D3k_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[R File](https://d3c33hcgiwev3.cloudfront.net/_0d2c90b40b32694737fc879a92bef9b8_R-script-Module-3.R?Expires=1703721600&Signature=LPxBTGHRuP~n5csmIjzwb-C~pGFsqqqMQwpv8jh8~BCrlx4zySXqzItPTepTuoVyzbeHmckidHs95XFUfkEWnfVds3hGy104ux~fl4x9BDfODWVMlRAPQwKwaTApK29JJPu7TPagiWZHniKnYKzoatpFyNsg-NEoWa74i9j5D3k_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[DATA\_3.01\_CREDIT](https://d396qusza40orc.cloudfront.net/phoenixassets/strategic-analytics-s12n/new/DATA_3.01_CREDIT.csv)

[DATA\_3.02\_HR2](https://d396qusza40orc.cloudfront.net/phoenixassets/strategic-analytics-s12n/new/DATA_3.02_HR2.csv)

# Wrap-up: identifying causes to effects

In this module, we have explained that one of the purposes of business

analytics was to understand the relationship between potential causes and

effects of interest.

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

But that basic correlation analysis was not satisfying:

it didn't allow us to consider the joint effects of different rivals all together.

And therefore, it didn't allow either to distinguish the most

important effects from the non-significant ones.

Play video starting at ::41 and follow transcript0:41

It may also lead to the wrong conclusions.

Play video starting at ::44 and follow transcript0:44

That's why you need to estimate statistical models to be

able to measure those relationships in a rigorous and quantitative way.

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

When approaching such models one can proceed in four steps.

One, assessing whether the resulting model is accurate.

Two, if the resulting model is accurate, what are the significant factors?

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

Three, for those factors, can we rank by order of importance?

Play video starting at :1:14 and follow transcript1:14

And then can we assess whether they are positive or negative for a business?

And finally, four, for the most important ones,

since they are part of a key message, we need to find a visual way to

report the relationship between the driver and the outcome we are interested in.

We've seen two examples.

How to model calls at the function of different financial and

social demographics factors.

Play video starting at :1:41 and follow transcript1:41

And how to understand what's driving the employee's attrition of a company.

Play video starting at :1:48 and follow transcript1:48

For each example, we spend some time to understand how

our results could be reported relevantly for a business audience.

Play video starting at :1:57 and follow transcript1:57

In particular,

we used model-free evidence of the relationships we unveiled.

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

Once the significance and the importance of a driver is confirmed with a robust

statistical method, we should rely on the visual evidences that do not need advanced

statistics to be understood, to build a shared presentation of the reality.

You can replicate all the examples that are presented in class thanks to

the recital, where we provided R scripts and

data to allow you to understand better how it works in practice.

<https://github.com/nicolasfguillaume/Strategic-Business-Analytics-with-R/blob/master/module3.md>

# Predictions & Forecasting: introduction

Many applications of business analytics aim at predicting future events.

Future customer behaviors, future outcomes, forecasting sales, or

predicting products, are all supply chain indicators and time series in general.

While managing a business you will try to anticipate what will happen,

when it will happen, and what will be the impact measured in a quantified way.

At the very least you need to have a good idea

of the range of the possible outcomes... and to obtain this objective

analysis is probably the most powerful tool you will have at your disposal.

In this module we will first explain why being able to forecast the future is

crucial, but we'll also discuss fallacies you want to avoid.

Once again, correlations may be a deceptive tool to predict the future.

As well known in statistics, causality and correlations are different things.

We'll cover different examples from value sectors.

Play video starting at :1:6 and follow transcript1:06

First, we'll take again the example of credit scoring and

HR Analytics, but now aiming at predicting what will happen.

Then we'll discuss a case of predictive maintenance using survival analysis.

Finally, we'll discuss seasonality in the context of the first example discussed in

this book using analytics for managing your supply chain and logistics better.

Play video starting at :1:32 and follow transcript1:32

As usual, all the techniques and examples presented during this video will

be replicable, thanks to the recital presented at the end of this module.

# Predicting events: sales, defaults, risks, churn, etc.

The ability to predict what is likely to happen in the future is probably one of

the most important issues for a manager.

Strategically, being able to anticipate the results of your actions

allows you to identify the efficient strategies to adapt.

Play video starting at ::26 and follow transcript0:26

Operationally, detecting events that will happen,

before they will happen, allows the operations to be more efficient as well.

At the end of the day, detecting patterns or

identifying causes and effects is interesting.

But what is really useful, hence actionable, is the ability to anticipate

the results of different actions you're considering to take.

Actually, one could even wonder whether you can say that you have any idea about

causality, if it's not directly related to your ability to predict events.

On the one hand, Clive Granger, a Nobel Prize winner,

actually claimed that causality between a cause and a consequence, could be tested

by the ability of the potential cause to predict the consequence.

Play video starting at :1:18 and follow transcript1:18

On the other hand, you should always be aware of the Post hoc,

Ergo Propter hoc fallacy.

It's not because event B happens after event A, that A causes B.

Post hoc, Ergo Propter hoc fallacies are very dangerous when looking for

actionable insights in practice.

The rooster sings each morning before the sun rises.

But if I killed the rooster, the sun will probably rise again.

Play video starting at :1:46 and follow transcript1:46

A business example could be taken from customer analytics.

If you see that there is a high correlation between customer behaviors,

for instance, when customers request a certain type of phone and

the fact that those customers are leaving the company afterwards.

We could think that we should prevent this behavior. Not necessarily,

because it may be a symptom and not a cause.

In this example, if this forms is for

instance a contract designation document, it's a symptom.

It's irrelevant to try to prevent it.

But we should, instead, see it as a signal to act.

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

Let's call those customers, proactively,

to understand how we could serve them better.

Play video starting at :2:30 and follow transcript2:30

Once again, we shouldn't jump to conclusions, but reflect on the actionable

recommendations to make using common sense and business acumen.

Predicting events has many applications in practice.

From prediction, are we targeting marketing, to sales and

product management or predictive maintenance.

There’s a plethora of usages and applications.

In this MOOC,

we’ll focus on three approaches that are widely used in the industry.

Play video starting at :3:2 and follow transcript3:02

One, how to predict which observations (custumers, employees, etc.)

are more likely to behave in a certain way in the future.

Play video starting at :3:11 and follow transcript3:11

That is, using classical classification techniques to predict events.

Play video starting at :3:17 and follow transcript3:17

Two, when is a certain event likely to happen precisely?

Play video starting at :3:22 and follow transcript3:22

And we'll be using survival analysis, on a predictive maintenance example.

Play video starting at :3:28 and follow transcript3:28

And three, how to model seasonal effects.

For instance,

when we have sales volumes that are varying from one season to the other.

At the end of this module, during the recital, we provide as usual our script,

allowing you to practice yourself the examples we've seen during the videos.

# Using classification and regression techniques to forecast

We can use the techniques,

seen during the previous module to identify the observations that

are the most likely to demonstrate certain behavior in the future.

Play video starting at ::18 and follow transcript0:18

Thanks to the classification regression techniques, we have built

a model relating causes and outcomes based on past data.

We can therefore plug in to this model

new data that was not used to estimate the parameters of the model.

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

That is what is called Out-Of-Sample-Data, and we can consequently predict

the expected outcomes for those new observations thanks to the original model.

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

First, let's take again the example of credit scoring.

We can estimate again the model based on the same data as what we did previously,

but then we can use the estimated model on new dataset - out of sample

that hasn't been used to estimate the model - and make predictions.

As you'll see during the recital, once we have estimated the model and

collected additional data,

we can use the function "predict" in R to apply this model to this new data set.

Play video starting at :1:21 and follow transcript1:21

In this particular case, since we know what should be the credit score,

we can compare the predictions with the actual values.

Play video starting at :1:30 and follow transcript1:30

We achieve a correlation of 98%. And

when we look at the results visually with a plots,

here with all the fitted values as a function of the actual ones,

we see that the predictions are indeed very close to what they should be.

Play video starting at :1:47 and follow transcript1:47

Now what could we do with this in practice?

Play video starting at :1:50 and follow transcript1:50

Many things actually.

We can identify the applicants who do not have credit score yet but

that should be acquired as clients.

Or the opposite: avoid including bad credits.

You see that once we have estimated the model,

we can predict the outcome of interest for any additional observation as

long as we have the explanatory variables used in the model.

Play video starting at :2:17 and follow transcript2:17

Let's now take the HR analytics example again.

It will be the last time we use it.

Play video starting at :2:23 and follow transcript2:23

We can apply exactly the same approach here.

We can use the predict function on the output of the GLM

function that was used to conduct a logistic regression.

Play video starting at :2:34 and follow transcript2:34

Here all the employees will be focusing on our analysis.

So, the prediction needs to be seen

as a probability that those employees will leave the company in the future.

Play video starting at :2:47 and follow transcript2:47

We then have for each employee in our sample a probability like this.

Play video starting at :2:54 and follow transcript2:54

What can we do with it?

We could, for instance, prioritize our actions and

if we do so we need to keep in mind that the probability informs us

on the employees that are most likely to leave.

But we also need to add to the analysis information about how much we

want to retain those employees.

As we've seen before in the dataset, we have information about the performance

of the employee thanks to the evaluation variable.

Play video starting at :3:25 and follow transcript3:25

Let's add it to the analysis and create a table with those pieces of information,

Play video starting at :3:31 and follow transcript3:31

probability to leave and performance.

We can then work on a 2x2 matrix.

Play video starting at :3:37 and follow transcript3:37

We first have the employees that are underperforming.

We should improve their performance.

Play video starting at :3:43 and follow transcript3:43

Then we have those who are performing okay, and that we should retain.

Among those we want to retain, some employees are not likely to leave soon.

So we should manage them as usual.

Play video starting at :3:56 and follow transcript3:56

And then, on the short run, we should focus on those with a good performance and

a high probability to leave, to assess the visual priorities for

each employee in a quantitative way.

We could for instance, multiply the probability to leave with the performance

Play video starting at :4:14 and follow transcript4:14

we then have a priority score

since the result of this product will be high for

the employees that we want to act on quickly, and low for the other ones.

Play video starting at :4:25 and follow transcript4:25

So if you rank our employees by this priority score,

we obtain something like this.

Play video starting at :4:32 and follow transcript4:32

The first column reporting the ID of the employee, the second, the probability

to leave, the third, the performance. And finally, the product of the latter two,

the priority score, which will allow us to prioritize our actions.

Now, we have a rank of the employees that we want to retain,

Play video starting at :4:51 and follow transcript4:51

the score being the combination of who is likely to leave and

how much we want to retain them.

Tomorrow, we can organize a face to face meeting with the first one - employee 9 2 8 -

then the second one - employee 5 8 8 - and so on.

Play video starting at :5:11 and follow transcript5:11

This type of prioritization is really useful in practice.

We could also have identified quick wins thanks to this approach:

employees that are high performance that were likely to leave very soon, but

now that we have identified them, we'll retain them.

This type of analysis is also used a lot to prioritize marketing actions

such as retention or in the case of propensity- to-buy or cross-sell and up-sell modeling.

Play video starting at :5:38 and follow transcript5:38

In such a context the probability interest

would be the probability that customer buys the product -

which can be estimated thanks to a similar model as what we've seen in the HR case.

And the equivalent of the performance variable could be the expected margin, for

instance, of this product.

But using a classification technique as a prediction tool has a major flaw.

Play video starting at :6:4 and follow transcript6:04

We may know whether an event is likely to happen in the future, but

we don't know when, exactly.

So let's now discuss techniques allowing to predict the expected time for

a certain event.

# Predicting when an event will happen with survival analysis

When interested in predicting when an event will happen,

one very often relies on survival analysis.

It's a set of techniques originally coming from life science.

As Keynes said, in the long run everybody dies.

But the pragmatic question is actually okay, but

how long will I enjoy life before it happens.

The answer is provided by survival analysis.

While it has many applications in life science,

it may easily be translated into business analytics.

Replace the event of "death" by another event and

you can apply to many different fields.

For instance we could model again how long before an employee leaves.

We now have identified what were the drivers of attrition and

who was likely to leave, but we don't know when exactly.

Since we have discussed HR analytics a lot already, we could use

another example instead and apply the same question to marketing analytics:

Play video starting at :1:10 and follow transcript1:10

how long before a customer churns?

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

From the modeling perspective, it's exactly the same question.

Play video starting at :1:18 and follow transcript1:18

Or even in grade scoring.

How long before a lender defaults.

Play video starting at :1:24 and follow transcript1:24

But let's take an example in predictive maintenance,

which is a topic we haven't discussed until now.

How long before a certain mechanical element breaks down, for instance.

Play video starting at :1:36 and follow transcript1:36

This question is actionable in practice.

Because if you can anticipate failures

you can allocate your maintenance resources more efficiently and

reduce the downtime of your processes by replacing the broken pieces proactively.

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

Let's take an example where we have statistics about 1,000 mechanical elements.

Play video starting at :2:1 and follow transcript2:01

Our dataset contains the lifetime of each element in weeks.

We report how long it has been used until now.

Or until it broke.

And whether it is broken with (a one) or still working (with a zero).

It indicates if the "event" we are interested in, here

the "death" of the element, happened already or not.

Play video starting at :2:24 and follow transcript2:24

Then, we have some measures about its environment.

A pressure index, a moisture index, and a temperature index.

We also have information about the team that is in charge of maintaining

each specific piece.

Team A, B, or C.

And finally, we know the provider of the element.

We can compute that we have 39.7% of broken pieces in the sample.

Those have been replaced already.

But that also means that more than 60% of them are still working well

since their installation.

One could think that we could rely on the standard linear regression

to find the causes leading to a failure.

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

If we do so, we obtain those results.

We see that we have only one or two variables that have a significant effect.

But actually, this is neither reliable nor accurate!

Because there are many elements that didn't break yet and

we don't know how long they will still last.

So for the 60% of the observations that didn't break until now,

we cannot really say when it will happen.

And since, with the standard linear regression,

we cannot estimate what's driving an event that didn't happen yet

we cannot say anything about those, and we cannot only include

these elements that broke either, because we would then have a biased model.

Play video starting at :3:49 and follow transcript3:49

This is actually, what we call in statistics, a right-censored problem.

We know who is "alive" until now, but even if we suppose that on the long run

everybody dies, we don't know when it will happen for those who are alive today.

Let me take an example of four employees who started to work with us at

the same time.

Let's say it's a month ago.

Play video starting at :4:14 and follow transcript4:14

Employee one left after a while.

We can hence measure the "time to death" for this employee.

The same for employee two.

Play video starting at :4:23 and follow transcript4:23

But employee three and four are still around.

It may be that in the long run, they leave,

but that part of the timeline - the hard part of this plot - is unknown.

It's censored.

That's why it's called a high-censored problem.

And so we cannot say what would be the time to death of employees three and

four in the end exactly.

But we -

we can make some assumptions.

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

Hence, we need to rely on a model adapted for right-censored situations.

A model that can estimate what's driving the probability to die or to break or

to default, even for the observations that didn't die yet.

Because the fact that they are still alive today still conveys some information.

So let's use the survival library in R.

As usual, I won't enter into the details of how these models work, and I'll just

focus on how we can interpret and use the result in a business-oriented way.

Play video starting at :5:25 and follow transcript5:25

We'll use the Survreg function,

which is basically a regression that adapted to a right-censored context.

We have to provide the lifetime and whether the element broke or

not for each observation as the dependent variable.

Play video starting at :5:41 and follow transcript5:41

And once we provide the pressure, moisture, temperature, team and

provider information as the exponential variables, we'll get these results.

Now in practice we should first access the quality of a model.

One could for

instance rely on an out of sample dataset, as we did with Credit Scoring.

To assess whether it does a good job or

not at fitting observations that were not used for the estimation of the model.

Play video starting at :6:11 and follow transcript6:11

Or we could only use data to a certain time point, for instance in the past,

and look at what's the quality of the predictions after the time point.

Play video starting at :6:22 and follow transcript6:22

But, each time, we'll been limited by this right-censoring issue.

We need to wait for

the death of an observation to know what is exactly its total time to death.

Play video starting at :6:34 and follow transcript6:34

You can do some tests in R yourself if you want, but you could see that

the model that we chose does, actually, a very good job at fitting the data we have.

Play video starting at :6:44 and follow transcript6:44

So let's assume that we trust the model here and go on.

As usual we identify the significant variables by looking at the p values.

The smaller the better.

Here we emphasize in bold the variables that have a p-value smaller than 0.05.

Play video starting at :7:3 and follow transcript7:03

Those are the moisture index, the temperature index,

whether the element is taken care of by Team C, and all the provider variables.

We then look at the sign of the estimates in order to assess whether the effect is

positive or negative for the expected lifetime.

Play video starting at :7:20 and follow transcript7:20

Positive means that the factor increases the chance of survival, and

negative means the opposite.

We see that moisture has, an average, a positive effect on the lifetime,

and that, everything else being equal, the units provided by providers two and

four have better expected lifetime than the others.

Now that doesn't mean that we should only work with providers two and four.

They may be more expensive or

they may provide elements that are used in conditions that are less demanding.

Play video starting at :7:55 and follow transcript7:55

But that should certainly be investigated.

Play video starting at :7:59 and follow transcript7:59

And the same for the negative effects, temperature, and

when team C is in charge of the maintenance.

It may be explained but we should investigate further.

And if there is no satisfying explanation, we should act upon it.

Now that we have a model, we can use its predicting power to anticipate failures

and do some predictive maintenance.

The predict function allows to use the result of the survival model estimations

for predicting the expected median "time to death" of each individual element.

Hence, for each observation, we can compare this expected time to death with

the current lifetime and compute the expected remaining lifetime, which

is just the difference between the actual lifetime and the expected time to death.

Play video starting at :8:50 and follow transcript8:50

Hence, we can now predict which elements

are the more likely to break in the near future.

Play video starting at :8:56 and follow transcript8:56

We obviously need to remove those that broke already.

Play video starting at :9: and follow transcript9:00

And if we then rank the observations by their expected "remaining time until death",

we then obtain this table.

It can consequently be used to prioritize our maintenance actions. And

replace the pieces that will break soon, before it happens,

by anticipating failures, acting proactively, and therefore decreasing

the downtime of the process, we benefit from more efficient resource allocation.

We avoid unnecessary tasks. Here maintenance of pieces that

still have a long expected lifetime. Which in turn, will make more

resources available to focus instead on actions that would be more impactful.

Play video starting at :9:44 and follow transcript9:44

That is, here, to replace elements that would have broken soon.

Play video starting at :9:48 and follow transcript9:48

In this video, we only provided an introduction to survival analysis.

We didn't enter into the details of different models available or

how to assess and compare their accuracy. But we saw that using

a library "out of the box" already did provide a lot of insightful information.

And at the end of the day, did the job very well.

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

Now, you should be careful in practice, because many hypotheses need to be

respected for the model conclusions and predictions to be valid and reliable.

But here, since we emphasize the managerial usefulness of those tools,

I focused on the interpretation and actionability of the standard approach.

If you like the topic, I would certainly advise you to learn more about

survival analysis on the web, in books and the like.

As I explain before, if you're just starting to deal with computer sciences

and statistics, you can see this training as an introduction to analytics and

the first step in your data science journey.

Play video starting at :10:52 and follow transcript10:52

In that case, you should go further in the understanding of those models.

Play video starting at :10:57 and follow transcript10:57

In contrast, if you're already a computer scientist or a statistician,

you can see this training as a way to work on all those analytics techniques

that you've learned in the past, but applied to a business context instead.

# Introduction to time series and seasonality

Now, sometimes we want to go one step further, and

we want to predict the whole time series of a certain type of event.

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

Typically, when managing the production of goods you want to be able to anticipate

what would be the sales, in order to match the production accordingly.

Let's retake the first example of this training,

where we had different SKUs with different volumes of expected sales and variabilities.

Anticipating the production was an issue.

Particularly for the wild bulls where we expected large sales but

we also knew that those sales could be highly variable.

Hence we say you should treat them on a case by case basis.

Let's do it and let's take a specific SKU to try and understand it better.

Play video starting at ::58 and follow transcript0:58

Now, let me unveil which product we have in front of us.

We are actually talking about a specific chocolate bar.

Play video starting at :1:6 and follow transcript1:06

As a chocolate bar manufacturer, it's important to be able to forecast the sales

accurately in order to allocate the right quantities of bars in store

proactively. On the one hand, should you have too much chocolate in store,

your inventory management won't be efficient.

You will waste a lot of money by placing products on the shelves,

that will not be purchased.

Without speaking of the products that passed the sell-by date and

will be discarded anyway...

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

On the other hand, if you underestimate the demand it may even be worse:

your turnover will be impacted negatively

because you failed to sell products you could have sold otherwise.

The problem is that the sales of chocolate bars typically vary over time.

You have certain periods where customers buy less and others when they buy more.

Play video starting at :2:3 and follow transcript2:03

Let's look at those sales over time.

We see that we have peaks and valleys.

The sales can go up to one thousand units in one month and

go very low during another one.

But we also see that there's some kind of regularity in the time service.

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

Maybe, we can identify patterns that could be used to build a model and

therefore improve our prediction ability.

If we succeed even though the sales are varying a lot

we could decrease our uncertainty about it to an acceptable level.

Play video starting at :2:35 and follow transcript2:35

We would then tame the wild bull.

Let's test it by using the month of the year as an exploratory variable,

we do a good job at modeling the sales.

Play video starting at :2:46 and follow transcript2:46

We can then estimate in R, what the effect of the month

using a simple similar regression where the sales would be the dependent variable.

And we see indeed, that many months have very strong effect on sales.

Remember that as soon as you see a star here it means that the variable

has significant effect on the outcome and

three stars means that the effect is strongly significant.

Play video starting at :3:15 and follow transcript3:15

We see that December, November and

February have on average a positive effect on sales.

A sign of the estimate is indeed positive in those cases and

also the months June to September report a decrease in sales on average.

It's not surprinsing if you are in a Northern Hemisphere Country,

those months are actually the hottest of the year.

Play video starting at :3:39 and follow transcript3:39

And people tend to buy less chocolate when it's hot.

Play video starting at :3:43 and follow transcript3:43

Appropriately with the coldest months of the year,

people will eat more chocolates then.

And if your are a country like France, for instance, it's not surprising that

people's consumption of chocolate peaks around Christmas, which is in December,

where we have the largest and positive estimated effect on sales.

Play video starting at :4:5 and follow transcript4:05

As usually we should find a visual way to report those results and

we could for instance use box plots as we do here

to report the distribution of sales per month.

A box plot allows for different cases here for different months

to report the distribution of a certain value here to sales.

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

It summarizes nicely the distribution of sales per month.

The box itself contains 50% of the values of the specific month and

the bar within the box reports the median value.

Play video starting at :4:40 and follow transcript4:40

As usual, here I present quickly this approach, but

I would advise you to investigate all the information a box report can report.

It's something really useful in practice.

Play video starting at :4:52 and follow transcript4:52

In any case, it allows (you) to compare in a very visual way, the differences between

the months and we see clearly indeed, which months report large sales.

And which months report smaller sales.

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

But does the model do a good job at fitting the variance we've seen before?

Play video starting at :5:12 and follow transcript5:12

Let's compare the sales as modeled by a regression and the actual values.

Play video starting at :5:18 and follow transcript5:18

We see that the fit is indeed very good.

We can therefore use the model we have to forecast what will be the sales for

each specific month in the future.

And while the sales- seen globally- are highly valuable variance

given a certain month, is a lot smaller than what we could have expected before.

Play video starting at :5:40 and follow transcript5:40

We reduced a lot the uncertainty, and we can now

use our ability to forcast future sales to allocate the stocks more efficiently.

We will maximise our sales,

while decreasing our cost of inventory management.

This was the last example for this module on predictions and forecasts.

You now have the opportunity to go over the recital once again.

# Recital M4 - Credit Score

Hi guys.

Welcome back for this modules recital.

This week we will add some protection and forecasting tools to our current tool box.

Play video starting at ::10 and follow transcript0:10

Last module, we tried to understand the relationship between causes and

consequences.

Play video starting at ::16 and follow transcript0:16

This week we're going to focus on using the knowledge build last week.

In particular, our ability to build classification and

regression models, in order to make predictions on out of sample data,

that is, data that the model has never seen before.

Play video starting at ::32 and follow transcript0:32

As usual, we'll review all the examples covered in class, and

explain how to use R to obtain the same outputs.

Let's get started with the credit score example we saw last module.

As we always do let's set our directory to the folder where we have downloaded

both the original credit scoring data set that we explored in the last module and

the new credit score and data set with the additional observation.

For me this is this folder and I set is as my working directory.

We then run this line in order to clean the memory of our current session.

And we then load the data set with the read.table function and

call the original data set dataold, and the new dataset,

datanew, with the new additional observations that the model

that we built last week has never seen before.

Play video starting at :1:26 and follow transcript1:26

Now if you do not remember the data set from last module, you should spend

some time exploring it again with the STR function and the summary function.

Let's go over the datanew data set very quickly.

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

If we run the STR function, we see that there are 100 observation in this data set

and 10 variables in this data set which are the same as is contained in the old

data set, namely the income variable, the rating variable, the cause variable,

the age variable, the education variable, the gender variable, the student variable.

And Married variable, the Ethnicity variable and the Balance variable.

Let's run the summary function on our datanew.

What can we learn on our variables in this dataset?

First, we see that the minimum income is $10.730.

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

The maximum is $182.73 in thousand U.S.

dollars and the average is $48.71 in thousands of U.S. dollars.

Now the rating variable has a minimum value of $112.

Play video starting at :2:45 and follow transcript2:45

A maximum value of $982 and

an average value of $375.4.

You should probably compare those numbers to the ones obtained in

the original data set.

In any case,

this gives us a good idea of the order of magnitude of the rating variable here.

Now, like we did last model, let's build a linear regression model

with the rating variable as our dependent variable and

all the other variables as independent variables.

Here we used a dot, and

the data we use in order to build the model is our dataold data set,

which is again the same data set that we used in the last model.

Let's run the line.

Now that we have a model that we call linreg and

they we assess how it performed as we did in the previous model.

We're ready to test the model on out of sample data.

What we want is the model to give us predicted credit scores for

the observations of our new dataset, the one that we called datanew.

Again, remember that this is data that the model has never seen before, NR.

In order to obtain predictions made by the model on new data,

we used the predict function.

The first argument is the model that you wanna use.

For us, it's linreg.

Then the new data argument is set equal to our new data

which in our case is the datanew as a data set.

Finally, a last argument allows us to choose the type of prediction

we wish to make.

Here we set type equal to response in close in order to

obtain the predicted credit score.

We then store the input in a new variable called predcreditscore.

Let's run the line.

Now let's go to the predcreditscore variable in order to see its output.

In order to do that you can highlight the name of the variable and

then press Cmd+Enter on the Mac and Ctrl+Enter on a PC.

Play video starting at :4:56 and follow transcript4:56

What the predict function has done is to predict a credit score for

everyone of 100 new observations and

it did so using the model dblock on the in sample data.

Now let's see how the model performs using correlation between fitted and

actual values on the dataold data set.

Play video starting at :5:19 and follow transcript5:19

As we did in the previous model, we use the core function.

Play video starting at :5:25 and follow transcript5:25

The fitted value is first and the actual value second.

We obtain a 98% correlation,

more precisely 0.9867324.

Now let's have a visual look at these results.

Play video starting at :5:45 and follow transcript5:45

By using a plot which has the fitted values on a vertical axis,

and the actual values on the horizontal axis.

Again, you see that the model does pretty well,

although not as well on the lower values.

Now that we know how the model does on in sample data,

let's check out how it goes on out of simple data.

Play video starting at :6:9 and follow transcript6:09

Here, again, we use correlation.

The first one here is our predictions and

then we have the actual values by typing datanew$rating.

Play video starting at :6:21 and follow transcript6:21

We get an an even stronger number of about 99%,

more precisely 9880, more precisely 0.9880.

Again let's see the output visually by using a plot.

Play video starting at :6:39 and follow transcript6:39

We have less data point cuz we only have 100, but

we see that the model does a pretty good job.

And here again, we see that for lower values it doesn't do as good of a job.

What's interesting here is that if we had more observation, or in this case, more

applicants for learned, for which we had values for all the explanatory variables.

We could give them a credit score and assess whether or

not it makes sense to grant them a loan, from a business point of view.

So that's it for this example and I'll see you in the next video.

# Recital M4 - HR example

Let's now go back to the HR example we explored in the last module.

If you remember, we wanted to see the relationship between a bunch of variables

in an outcome, namely whether the employee left the company or not.

Play video starting at ::14 and follow transcript0:14

To do so, we built a logistic regression model,

which performed pretty well on the in sample data.

Where we're going now is to use the model

on the employees who are currently still with us to see the probability

that they are leaving in order to undertake actions aimed at retaining them.

As always let's set our current working directory to the folder where have

downloaded the original HR data set that we explored in the last module.

And the new HR data set with the additional observations.

Play video starting at ::51 and follow transcript0:51

Once this is done,

we run this line to clean up the memory of our current R session and

we are now ready to load both of our data sets with the read.table function.

You already know the argument, the header equal true, and

the separator equals comma, my input.

So let's load the data, old data set,

which is the same that we used in the last module.

Play video starting at :1:21 and follow transcript1:21

And the data, new data set, which contains only

the employees who are still currently in the company.

Like in the previous video,

if you do not remember the data set from last module, you should spend time

exploring it again with the STR function and with the summary function.

Let's go over quickly the data, new data set,

in order to get more familiar with it.

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

There are a thousand observations in this data set and the six variables that you

already had in the data set that we used last module, the data, old data set.

Play video starting at :2:3 and follow transcript2:03

So we have the satisfaction level,

the last project evaluation, the number of project worked on in the last 12 month,

the average number of monthly hours, the time spent in the company.

And whether or not the employee had a new born within the last 12 months.

If we now run the summary function on data new, what do we find out?

First we find out that about 20% of our employees had a newborn within the last 12

months, so that might be something that we want to keep in mind when

designing employee's perks or working on an employee's well being.

Then we find out that on average our employees worked 100.2 hours per month.

And that they have spent on average,

a little bit over three years in the company.

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

Let's now rebuild the logistic regression model that we built in the last module.

You remember that in order to do so we use the GLM function and our first

argument is dependent variable which here is the left variable then tilt it.

And then the dot indicates that we want all the other variables to be used

as independent variables.

We then set family argument equal to binomial and in parentheses, logits.

And the data that we want to use in order to build the model and

compute the coefficients is the original data set, which is data old.

Play video starting at :3:33 and follow transcript3:33

We already assessed that the model was performing well, so

we can move on to using it to make predictions for out-of-sample data.

Like in the previous example, we are going to use the predict function and

again our first argument is the model that we call lugreg,

the second is the new data that we want to use the model on.

And in our case it's data new.

And then the type of predictions that we want to make and in our case,

we set it equal to response in quotes.

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

Now that we have stored the output of the predict function in PROBATOLEAVE,

let's go PROBATOLEAVE in order to see the output.

Play video starting at :4:17 and follow transcript4:17

What we see is that for each of our 1,000 observations, we get a probability.

Let's now organize the output nicely in a data frame by using

the data.frame function on PROBATOLEAVE.

And we store the output in predattrition.

Play video starting at :4:36 and follow transcript4:36

Let's see the data frame by using the view function.

Play video starting at :4:41 and follow transcript4:41

Now, we get the same thing but in a nice table which will be easier to use.

Play video starting at :4:46 and follow transcript4:46

Let's go back to the script.

At this point we can use the information in a lot of different ways.

Play video starting at :4:52 and follow transcript4:52

But let's focus like you did in class on using it to

prioritize our efforts to our best performers.

So the first thing that we want to do is to add the performance variable to

our predattrition data front.

To do so, you can type predattrition dollar sign performance.

And you set it equal to data new dollar sign LPE,

which is the last project evaluation and a fairly good measure of performance.

Play video starting at :5:21 and follow transcript5:21

Now let's view the output with the view function.

Play video starting at :5:26 and follow transcript5:26

You can see now that for

the first observation, there is a probability and a performance.

Let's now go back to our script, we can build a plot with the performance

on the vertical axis and the attrition probability on the horizontal axis.

This is what we get.

If we wanted to make the output more visual, we could add vertical and

horizontal lines with the AB line function like we did before.

And maybe add some text labels using the text function.

And this is what I'm going to do with the code that I'm going to copy paste here.

You can explore the code here on your own, but essentially,

we can see here very visually that these are the employees that we want to retain.

Play video starting at :6:11 and follow transcript6:11

Their likelihood to leave is fairly high.

That is the probability to leave and

their performance level is also pretty high.

So we definitely wanna target our actions to these people.

Now can you think of a way of prioritizing among those employees

in a much easier way?

Play video starting at :6:34 and follow transcript6:34

As explained in class we could establish a priority score by

multiplying performance and the probability to leave for each employee.

So we can add a priority column like you see here to

our predattrition data frame and the priority would be

the multiplication of the performance for each of the employees,

each observation and their probability to leave.

Let's run the line.

Play video starting at :7:6 and follow transcript7:06

Now if we view the output with the view function, this is what we get.

For each observation we have their probability to leave,

their performance, and then their score which is a result

of the multiplication of their probability to leave and their performance.

Play video starting at :7:24 and follow transcript7:24

Now let's turn back to our script in order to make our work a little bit easier.

What we can do is to order the output in decreasing order.

Play video starting at :7:35 and follow transcript7:35

And you can do that using the order function and

setting the decreasing argument equal to true.

So what we're doing here is that we're ordering from the employee with

the highest priority score to the one with the lowest priority score.

Because we're ordering by priority scores.

Now to do so we create a new data frame called Order Predattrition and

we run the line that we just explained here.

Play video starting at :8:6 and follow transcript8:06

The order pred attrition data frame contained all the data from predattrition

but the organized in decreasing order of the priority scores thanks to the call to

the order function.

Now let's check out the output with the view function

Play video starting at :8:23 and follow transcript8:23

called on the order predattrition data frame.

Play video starting at :8:28 and follow transcript8:28

And now we see that the employee with the highest priority is ranked first,

so based on our resources now we'll be able to decide

which employees we want to target our efforts on in this order.

Maybe it'll be the first two, maybe the first five, maybe the first 20,

depending on the company and the resource that we have.

But this gives us a really nice way to prioritize our efforts.

So that's it for this example, and I will see you in the next video.

# Recital M4 - Predictive maintenance example

Welcome back for this third example.

In this video we're going to go over the predictive maintenance example that you

explored in class.

As always, we'll focus on learning how to perform a survival analysis in

arc in order to obtain the same results shown to you during the lectures and

most importantly for

you to be able to perform your own survival analysis on your own datasets.

Bear in mind that our issue in this example is to estimate the remaining

lifetime of our PCs, so we can better organize our maintenance efforts.

All right, let's get started.

We set our working directory to the folder where we have

downloaded the predictive maintenance dataset, and

we then run this line in order to clean up the memory of our current arc session.

Play video starting at ::42 and follow transcript0:42

We're now ready to load our maintenance dataset and coiled data.

And we load it using the read.table function using the set argument and

the header argument.

Now, let's get familiar with our data by calling the str function.

Play video starting at ::58 and follow transcript0:58

And we find out that we have 1,000 observations in this dataset, and

seven variables.

Namely, the livetime variable,

which reports in weeks, how long it has been used until now.

Then the broken variable, which takes the value 1 if the piece is broken,

and 0 otherwise.

Then the pressure index, the moisture index, the temperature index,

the team that is in charge of maintenance, and the provider that supplied the piece.

Let's now take out some summary stats on our dataset to understand the order of

magnitude of our variable, and getting better integration of our data.

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

First thing that is of interest is the proportion

of the PCs of observation that are broken.

And we see that we have 0.397.

So about 40% of the PCs in this dataset are broken.

And then, we may be interested to know that the mean

pressure is 98.6, mean moisture index is 99.38,

and mean temperature index is 100.63.

The last thing that may be of interest,

is to look at the breakdown of PCs maintenance allocation to teams.

336 PCs are allocated to team one, 356 to team B, and 380 team C.

One thing that we've done before in order to predict the value of

a numerical variable is to build a linear regression model.

So let's try it out here.

Let's call our model lin reg model and we use the ln function.

Our dependent variable is lifetime and

then the dot indicates that we use all the other variables and then

the minus broken indicates that we use all the variables, minus the broken variable.

But we could add all those variables to the model after the tilde and

separate each of them by a plus sign.

But again, it is just more efficient.

And I want you to know that you can do that, if you want to.

And then we set data equal to data to use our data dataset.

Let's run the line.

Now, let's check out the output of our model.

What we see is that It looks like the fact that team C

which was in charge of the maintenance of the piece is statistically significant,

as well as the fact that the piece was provided by provider number three.

And both of them seem to impact negatively

on the lifetime if we look at the co efficient and its sign.

But as was explained in the lecture,

using the output of a linear regression, this case is inaccurate.

In this case,

we cannot rely on the output of the linear regression model for a very simple reason.

For 60% of the observations, those are not broken, the lifetime is not

the live time until the PC was broken, but rather, the lifetime until now.

So the truth is, we don't really know what the final value of lifetime will be for

PCs that are not broken yet.

All we know is that they are still working until now.

As Professor Glady explained in the lecture, this is a right-censored problem

that we will tackle in this video by using a survival model.

In order to create such a model in R, we need to install the survival package, and

in order to do so, we need to use the install.packages function,

open parenthesis, and survival in quotes.

Play video starting at :4:20 and follow transcript4:20

We read the line, and it may take a few seconds.

Now we need to load the survival package by using the library function,

open parenthesis, and survival.

Without quotes this time, we run it, and the package is loaded.

Now, please note that while you will not have to reinstall the survival package for

your subsequent analysis,

you will need to load the package every time you start a new session.

In any case, we're not ready to start our survival analysis.

Our first step is to set our dependent variables and or

we can do that by using the Surv function, and in this case,

we want to provide the lifetime and whether or not the PC is broken.

And we call the output dependentvars.

We're now ready to build a model that we will call servreg.

To do so, we use the servreg function.

We input our dependant vars, which is the output of the previous line,

tilde, and then our independent variable, namely pressure index,

Play video starting at :5:22 and follow transcript5:22

moisture, temperature index, team provider.

We then set the dist argument equal to Gaussian,

in quotes, because we want distances to be computed using the Gaussian method.

And as usual, we set data equal to data, which is our dataset.

We then run the line.

At this point, we should assess our model on out of sample data for example.

We won't do it here because we checked it for you.

But keep in mind that you should not use a model without being confident that it

does well, or at least without knowing its limits.

Let's now check the output of the model.

As we always do,

let's check out the p value to see which variable is statistically significant.

We see that first, the moisture index is statistically significant.

The temperature index as well.

If the PC was maintained by team C and if the PC was

provided by provider two, three or four are all statistically significant.

Now, to assess whether the effect is positive or

negative on the expected lifetime, we look at this sign of the coefficient.

For our moisture index it's positive.

For temperature index it's negative.

For Team C it's negative.

For Provider two It's positive.

And for provider three it's negative and provider four, it's positive.

When I say positive or negative,

it is on the expected lifetime that I'm talking about.

So we see that if the team was supplied by provider two or

provider four, the expected lifetime is positively impacted.

But in any case, it is information that is worth knowing, and

that is worth investigating.

Play video starting at :7:3 and follow transcript7:03

Now that we have a working model, let's go back to our initial problem, which was to

estimate the remaining lifetime of our PCs, which are not currently broken.

To do so, we're going to make predictions using the predict function,

which will output the expected median of each individual element.

Our first argument is the model that we just built and

that we're going to use to do our predictions.

Our new data argument is set equal to data which shows that our dataset.

And we set the type argument equal to quantile, in quotes, and

P for percentile, equal to 0.5 in order to get the expected median time today.

And we store the output in EBREAK.

Now to make an interesting report, we can organize the output

of EBREAK in a data frame, with the data frame function.

And we call the output Forecast.

We can then add a column with the lifetime for

each observation by typing Forecast$lifetime=data Data$lifetime,

and then a column indicating whether or not the piece is currently broken

with forecast$broken = data$broken.

And the last column, which we will call remaining LT, for remaining lifetime,

which computes the remaining lifetime by subtracting the lifetime variable from

the original dataset to EBREAK, which is the output of our predict function.

Let's check out the output by typing View.

Play video starting at :8:46 and follow transcript8:46

Now the output is interesting, but we see that we still have the PCs that

are broken, and that we're not really interested in,

since it's too late to do some maintenance on those.

And the other thing is that our output is currently not prioritized.

We don't have the PCs that will be broken shortly that we should focus on.

What can we do to improve the output?

First, we can reorder our forecast data frame and

order it by increasing remaining lifetime by using the order function.

And from forecast, we can select only the PCs that are not broken yet.

That is those for which the broken variable is equal to zero.

Play video starting at :9:29 and follow transcript9:29

Now let's take out the output of ActionsPriority,

which is our final dataset.

Play video starting at :9:36 and follow transcript9:36

That's much better.

We can now allocate our maintenance staff to the PCs most likely to break soon,

and maybe avoid some breakdown timeout.

So that's it for the third example.

I will see you in the next video.

# Recital M4 - Chocolate Sales example

All right guys, let's get started with our last example for this module,

it's going to be yummy, and it deals with the chocolate sales forecasting example.

Play video starting at ::11 and follow transcript0:11

As always, let's get started by setting our directory to the folder where we've

downloaded the chocolate data set.

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

Once it is done, we clean up the memory of our current R session.

And we are ready to load our data set with the read.table function and

the usual arguments that we use.

Now to get familiar with our data set, we use the STR function.

Play video starting at ::41 and follow transcript0:41

And we call it on the dataset that we've named data.

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

What we find out is that we have 120 observations in this dataset and

four variables.

Namely the time variable which goes from one to

120 and which essentially means that we have ten years of monthly data.

Then the cells variable which corresponds to cells in thousands of units.

Then the year variable which tells us the year of the particular data point.

Play video starting at :1:15 and follow transcript1:15

And the month variable which tells us the month of the particular data point.

Well, until now we've always called this summary function on the entire data set.

It does not really make sense to have the mean value for the time.

Or for the year, or the month variables.

So, what we're going to do here is to call the summary function

on the unshelf by typing

summary(data$sales), and

we find out that the minimum sale was about 37,000 units,

the maximum was over a million units.

On average 216,000 units were sold.

Let's plot the sales as a function of time to get a better of our data.

To do so we use the plot function.

Time goes first because it will be on the horizontal axis and

cells second because it will be on the vertical axis.

Then we add a main title with the main argument and

access levels with x lab and y lab.

And we use the ylim argument in order to set a specific limit to

the y axis wider than the default, so we multiply the max,

the maximum value of cells by 1.2.

And our last argument here is the type of loading that we want to do.

And we said if equal to l in quotes, which is lined.

Let's run the line.

What do we see?

We see that the data shows seasonality.

Play video starting at :3:6 and follow transcript3:06

And most importantly, that is regular over time.

It looks like there is a peak somewhere around, let's say the 12 months,

and here again at, it looks like it could be month 24.

And so on, and so on.

Now as you did in the lecture, let's build a simple linear regression model,

with the sales as the dependent variable, and month as the independent variable.

And then the data that we use to build the model is our data set,

which is called data.

Play video starting at :3:45 and follow transcript3:45

If we take the output of the model, what do we see?

We find out that the most statistically significant predictors of sales

are February, August, November and December with three stars,

and then June, July, and September with two stars.

You remember that in the star system relies on the p value and for

a p value below .05,

which is usually deemed statistically significant, you get a dot.

Below .01, you get one star.

Below .001 you get two stars.

And then the closer it gets to zero you get three

which is extremely statistically significant.

Now if we look at the T value and the sine of the coefficient

we find that December followed by November and

February have the strongest positive effect on sale.

And as explained by Professor this mostly makes sense since a lot of people in

the Northern hemisphere buy chocolate for the end of the year holidays.

Now, another interesting information here would

be to see the distribution of sales for each month.

To see that information, you can create a box plot.

Because the month variable is a factor,

R will automatically represent the sales per month using box plots.

Play video starting at :5:17 and follow transcript5:17

Like this, we then add a main title, access labels,

and use the XM argument like we did before.

And we get this plot, which shows you, for each month,

the distribution of all the sales that happened in January, no matter the year.

Again, for February, March, April etc., etc.

Now BOXPLOT is a very interesting visual way to understand if there's

a lot of variation in the cells of a particular month.

For example, in December you see that there is a lot of variation,

while in July the data is very similar from year to year.

Now how does the model perform on past data?

Let's plot the actual sales as a function of time and

add the sales as provided by the model.

Play video starting at :6:8 and follow transcript6:08

To do so, we use the plot function with time on the horizontal axis,

and sales on the vertical axis.

And again we add titles and access label that you can explore in your own time.

Play video starting at :6:28 and follow transcript6:28

In order to add the fitted values to the same plot we use the lines function

were time is on the horizontal access and the fitted values on the vertical axis.

Then we set the type argument equal to l in quotes because we want a line.

The col argument equal to blue in quotes because we want this line to be blue.

And the lty argument allows you to specify

the type of line that you want to use, and 2, allows you to have a dashed line.

Let's run the line.

Play video starting at :7:5 and follow transcript7:05

Now we see that the model does really well and

it also allows you to reduce a lot the insurgency for stock management.

Now to make your plot more explicit,

it would be a good idea to add a legend by using the legend function.

The first argument here said equal to top left in quotes,

lets R know that you want your legend to be at the top left corner of the plot.

We then use the C function to set the text that we want to use in

order to describe our line.

Then again in order to describe the type of lines that we want to use.

And we maintain the same order for all the arguments.

So the actual sales have an lty argument set to 1 because it's a complete line.

While the sales by the model have an lty argument

set to 2 because it's a dashed line.

And then again for the col argument we set the actual sales

equal to black and the sales by the model equal to blue.

Let's run this line.

Play video starting at :8:16 and follow transcript8:16

We do have a legend here.

Play video starting at :8:18 and follow transcript8:18

That's it for our fourth and last example of this module.

I hope you had a good time and learned a lot.

And now it's time to go and eat some chocolate for real.

# Script and dataset files to replicate recitals

[R script module 4](https://d396qusza40orc.cloudfront.net/phoenixassets/strategic-analytics-s12n/new/FSBA_SCRIPT_MODULE4.R)

[DATA\_3.01\_CREDIT](https://d396qusza40orc.cloudfront.net/phoenixassets/strategic-analytics-s12n/new/DATA_3.01_CREDIT.csv)

[DATA\_4.01\_CREDIT2](https://d396qusza40orc.cloudfront.net/phoenixassets/strategic-analytics-s12n/DATA_4.01_CREDIT2.csv)

[DATA\_3.02\_HR2](https://d396qusza40orc.cloudfront.net/phoenixassets/strategic-analytics-s12n/new/DATA_3.02_HR2.csv)

[DATA\_4.02\_HR3](https://d396qusza40orc.cloudfront.net/phoenixassets/strategic-analytics-s12n/new/DATA_4.02_HR3.csv)

[DATA\_4.03\_MNT](https://d396qusza40orc.cloudfront.net/phoenixassets/strategic-analytics-s12n/new/DATA_4.03_MNT.csv)

[DATA\_4.04\_CHOC](https://d396qusza40orc.cloudfront.net/phoenixassets/strategic-analytics-s12n/new/DATA_4.04_CHOC.csv)

# Wrap-up: forecasting events

In this module, we have explained how predictive models can be used to

anticipate future events, prioritize our actions and act accordingly.

Play video starting at ::20 and follow transcript0:20

We first presented how descriptive models - such as a classification or

regression tools - can also be used to forecast future events.

Actually, we even pointed out that if those models were not

able to predict what would happen, it meant that they

actually did a poor job at describing the reality of the situation in any case.

Play video starting at ::43 and follow transcript0:43

In business, if you want to understand something better,

it's mainly to know how to act more efficiently.

Therefore, if an insight cannot be used for

anticipating the results of future actions, it's not, technically speaking,

actionable, and it's then irrelevant from a strategic business analytics perspective.

We discussed three examples in this fashion.

We took the credit scoring example. This time, to estimate what should be

the credit score of an individual based on past data and predictors.

We also took once again the HR analytics example to predict

who were the employees that were the most likely to leave.

And we came back to the very first example of this training, the supply chain and

logistics management of SKUs, analyzing expected sales and

their variances to approach the problem of seasonality.

We also discussed the right-censored example.

That is when the event to happen hasn't necessarily happened already.

Which actually is the case in many real-life applications

such as positive maintenance, as we have seen.

We then used survival analysis to design a robust and reliable model.

Play video starting at :2:3 and follow transcript2:03

Each time, we demonstrate how accurate FORECASTS can be used to act,

prioritize, or be more efficient.

As usual, for each example, we spent some time to understand how our results

can be reported relevantly and visually for a business audience.

Play video starting at :2:22 and follow transcript2:22

You can replicate all the examples that are presented in the videos,

thanks to the recital, where we provided R scripts and

data to allow you to understand better how it works in practice.

(Required)

English

​

Help Us Translate

# Reporting your results: introduction

During this MOOC we've introduced the most important techniques used in business

analytics.

In each case we've discussed how to present the results in a visual,

relevant and insightful way, but that's not the end of it.

Once you've produced the most important summary statistics and

prepared them in a business oriented and communicable way,

you also need to present it to different audiences in practice.

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

It may be to your management, your client, your teams, or different stakeholders,

but it almost always needs to be communicated to someone in any case.

Play video starting at ::50 and follow transcript0:50

In this module, we'll discuss how you can present

your business analytics work to a business audience.

We'll explain that you need to find an angle and tell a story.

You should absolutely avoid giving the impression that you're presenting

a list of results that are not connected to each other.

Instead, you should take the audience by the hand and

stick to the recommendations you want to conclude on.

Play video starting at :1:19 and follow transcript1:19

We'll discuss how to structure your story and your slides, and

also present some of the most used visualization tips and tricks.

Play video starting at :1:29 and follow transcript1:29

As usual, there will be a recital at the end of the module, so

that you can start practicing by yourself.

Now, since it's all about the story, it's also a matter of style.

You'll certainly see different styles for structuring presentations, stories or

slides.

But while I would agree that there is no single way to tell a story,

there are certainly poor ways to do it.

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

If one may argue with some of the choices I'll make in terms of story telling and

presentation, there are certainly pitfalls that you should absolutely avoid.

Play video starting at :2:3 and follow transcript2:03

And let me explain what they are.

Play video starting at :2:6 and follow transcript2:06

Here, I'll focus on a certain style -

quite consulting oriented - and much "expertise" driven.

Because that's very often with Strategic Business Analytics

experts are hired for.

But you can easily adapt those pointers to any other type of presentation.

Let's start.

(Required)

English

​

Help Us Translate

# It's all about the story

Remember the best presentation you've seen and the worse.

You'll notice that the best presentations always share characteristics.

Good presentations are very often story driven.

The presenter tells a story that you want to follow.

Play video starting at ::26 and follow transcript0:26

In contrast, poor and boring presentations are not.

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

For instance, particularly when dealing with technical topics,

if you are presented with a list of results without any relationship whatsoever,

or the presenter starts focusing on technical details that nobody cares about,

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

they are likely stop paying attention.

When you'll be presenting the results of your own analysis, it's tempting to forget

that you're not talking or hearing your own voice but to pass a message, and

accordingly you shouldn't focus on what you want to say about your work:

for instance, how technical it was or how difficult some of the analyses were,

but focus on the key messages you want to convey to your audience.

Play video starting at :1:14 and follow transcript1:14

And so, a good business analytics story has a good beginning.

And I'm not talking about an opening joke.

Even if that could work, I'm talking about starting your presentation

with a relevent, to the point, hookup.

For instance, there is no need to start lingering

about the history of the company or its sector in general.

Play video starting at :1:36 and follow transcript1:36

Imagine that you're talking to your client as a consultant.

Play video starting at :1:41 and follow transcript1:41

He probably knows his own company better than you.

Instead, you should directly identify the issue at hand in a quantitative and

visual way to convince your audience that what you will be presenting

will be worth their time.

Play video starting at :1:56 and follow transcript1:56

For instance, what's the pain point you want to solve?

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

Are you trying to induce employees' attrition?

Say it directly and give us a number indicating how serious the problem is.

Even better, show us a graph about the number of people leaving each year.

Are you trying to serve your customer better?

Why is it important? Is it because a recent satisfaction survey was terrible?

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

Are we losing market shares?

Use quantitative information as soon as possible to demonstrate

the relevance of what you will address later in the presentation.

It may be a pain point: something to solve,

or it may be an opportunity to take.

But it cannot look like you're playing with data for the fun of it.

Play video starting at :2:46 and follow transcript2:46

If there is a reason why we're listening to you,

we should know it within the first 30 seconds, and

by the way, don't forget to address it by the end of your presentation.

I very often see presentations that claim that they would address a first issue and

we're actually addressing another one as the story progresses.

A great storyteller always delivers what was promised or foreshadowed.

Play video starting at :3:14 and follow transcript3:14

Now as a matter of fact, in practice, the key issue of the presentation is

very often discovered the other way around. The pain point or opportunity that you

will be presenting at the very beginning of your presentation will usually only be

identified within the course of your analysis after you've concluded your analysis!

Play video starting at :3:35 and follow transcript3:35

What happens is that while analyzing the data

you will discover opportunities to take or problems to solve only after a while.

Play video starting at :3:44 and follow transcript3:44

But when you present your work, it shouldn't follow the order of

the different actions you've taken in the course of your analysis!

It should be restructured after you've come to your conclusions

to ensure that everything you do from the beginning to the end looks like it

has been aimed at addressing the issue raised in the introduction.

Even though, when you've done your analysis,

the issue was only identified AFTER your conclusions.

You should approach the structure of your presentation as a storyteller.

It's the story you want to tell that drives the structure of the presentation,

not the way you've analyzed the data. And when you 're presenting your conclusions and

your recommendations, don't forget that a manager has limited resource and time.

At the end of the day, she can only focus on what's expected

to have the maximum impact on the smallest effort.

Play video starting at :4:44 and follow transcript4:44

Remember what we said about the need for efficiency in most of the modules.

Keep it in mind when you're writing your own recommendations.

You need to focus on the most efficient actions to take.

So once you've understood what are the actions that could solve the issue

you've raised,

you need to make sure to prioritise those actions according to their efficiency.

Play video starting at :5:10 and follow transcript5:10

Do not make an unordered list of all the actions that could be taken.

Present the quick wins or the low hanging fruit first.

Those actions that could be taken easily, because needing little effort,

but that would have a huge positive impact on your business.

Play video starting at :5:30 and follow transcript5:30

And only then explain what should be done

in the long run that will be impactful as well.

But that may require more effort and at the end you'll probably

decide to discard the less impactful or unrealistic recommendations in any case.

Play video starting at :5:48 and follow transcript5:48

If it doesn't sound relevant or

easy to do on paper, it will probably be worse in practice.

Play video starting at :5:55 and follow transcript5:55

And so you need to connect logically the issue that you've raised to

the conclusions that you've drawn.

Since we're dealing with business analytics you need to explain

why you need the data you've collected to address the issue, and

why the methodology you've chosen is relevant.

That doesn't mean that you need to enter into all the little gory details of your

approach, but it means that your presentation should be self-contained.

It should contain everything needed to understand the why, the what,

and the whole of your results.

Play video starting at :6:31 and follow transcript6:31

I very often do the test of the executive assistant.

Play video starting at :6:35 and follow transcript6:35

Does a smart executive assistant understand your analysis?

Play video starting at :6:40 and follow transcript6:40

Someone who knows the business, but is not a computer scientist or statistician.

Play video starting at :6:46 and follow transcript6:46

If the answer is no, work on your presentation until it's clear enough and

the story flows well.

Remember something, the point is not to show how smart you are and

how advanced your methodology is.

The point is to obtain the buy-in from your audiance.

And nobody likes to feel stupid or to listen to the bragging of the first of the class.

Be as clear and to the point as possible.

It's in your own interests.

In the same fashion,

do not start explaining in detail why something didn't work.

Play video starting at :7:21 and follow transcript7:21

Very often,in the course of your work, you will start conducting analyses that will be

fruitless! It's very frustrating.

Play video starting at :7:28 and follow transcript7:28

And it's very tempting to explain at length afterwards that you've worked for

so many hours on something that has been shown to be irrelevant in the end.

Sometimes, if it may indeed be a question raised by the audience but is not really

relevant for your conclusions, you may decide to keep it in (the) appendix.

That's fine.

But if it's not really relevant for your story, discard it and move on.

Just say "one could have made this analysis, we did it and

it didn't work because of X" and then go on with the rest of your storyline.

I know it will be difficult to let go because you suffered on the task, but

unfortunately nobody else than you cares.

So don't spoil your story with details that are irrelevant for your audience.

Play video starting at :8:16 and follow transcript8:16

And that is actually the question that should drive everything else.

Play video starting at :8:21 and follow transcript8:21

What is relevant for your audience and what's not?

Play video starting at :8:25 and follow transcript8:25

And that depends on who is in front of you.

So who is your audience?

It is an executive?

Play video starting at :8:31 and follow transcript8:31

Be very concise with one or two executive summary slides

emphasizing what should be done and what will be the expected benefit.

Play video starting at :8:41 and follow transcript8:41

Is it a technical person?

Play video starting at :8:43 and follow transcript8:43

Then you should take the opposite direction and

you may want to describe the methodology in detail.

Play video starting at :8:50 and follow transcript8:50

The same goes for the style.

Some people like that you start from the big picture and

zoom into the details on the afterwards.

And some people, instead, like that you do the opposite.

Build your story piece by piece and

present the overall conclusions only at the end.

To wrap up on the storytelling, let's say that your presentation

shouldn't look like a list or a series of analysis conducted separately.

Play video starting at :9:17 and follow transcript9:17

A bunch of interesting facts, but not really related whatsoever.

It's a story.So start first with the pain point to solve or an opportunity to take

Think of a good movie you've seen recently.

It probably starts with an average Joe who is suddenly in front of a problem.

Play video starting at :9:36 and follow transcript9:36

You should have the same type of hook up.

Play video starting at :9:39 and follow transcript9:39

Then explain how you approached it.

Describe the data, but only what's relevant for your analysis.

Do not waste too much time on what didn't work and

focus on having a flow of insights that follow each other and make sense.

Play video starting at :9:57 and follow transcript9:57

You want to attain your conclusions logically and

propose recommendations that will look like there is almost

no other common sense actions to take than those that you propose.

Play video starting at :10:11 and follow transcript10:11

And the way you will tell your story will depend on your audience.

Play video starting at :10:14 and follow transcript10:14

What's relevent for them and what's not?

Play video starting at :10:18 and follow transcript10:18

Some rationales will resonate better to a certain type of audience

Play video starting at :10:25 and follow transcript10:25

while others won’t. If you do your homework and

adapt to your audience by telling your story, most of the stakeholders

will buy-in your recommendations by the end of your presentation.

What you want to obtain is a shared representation of your reality

between you and your audience.

This is a key success factor in getting people to move in the same direction.

Play video starting at :10:49 and follow transcript10:49

And it will work even better if they feel they participate in the reflection and

come to the right conclusions by themselves.

# One slide / One idea

Now that we discussed the articulation between the parts of the presentation,

let's discuss the structure of the slides themselves..

The basic idea is that you should see your slides as a way to summarize as much

information as possible, while keeping the right balance of how clear they are.

The first thing I need to say about slides,

is that there are several approaches one could take.

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

One of the most popular approaches when making a presentation,

is to see your slides as a support to your oral presentation.

In that case, a slide should have as little text as possible.

Play video starting at ::49 and follow transcript0:49

It would just be an image, or at least something very visual.

Play video starting at ::53 and follow transcript0:53

Here we'll take exactly the opposite approach.

We'll see a presentation as a written document that is not necessarily

a support to an oral presentation,

but (is) instead self-contained and presents first and

foremost information that stands alone.

Play video starting at :1:12 and follow transcript1:12

It might also be used for an oral presentation afterwards if needed.

But most of the time, as a matter of fact, it’s not.

This is a typical consultant approach,

where your documents will be passed around teams of different persons.

And you will not necessarily have the opportunity to defend it.

Play video starting at :1:31 and follow transcript1:31

And sometimes it may happen that you will have this opportunity

to explain what you've produced in a face-to-face meeting.

So it's finding the right balance between several objectives

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

and to attain those objectives all at once, you should structure.

a slide as something that can be read at different levels.

People will sometimes only read the title

and that's why your title needs to be informative.

Instead of entitling your slide, "conclusions summary", for

instance, replace it by the actual conclusions of that slide.

Play video starting at :2:8 and follow transcript2:08

An example could be, "should we adapt or go to market to focus on our core business?"

This title in this case is an informative call to action.

It's longer, but it's useful, it's what we call an action title.

What's also nice is that if you do that, and then read all the titles from

the first line to the end, you should have the outline of your story.

It's a neat way to check that everything holds together.

Play video starting at :2:37 and follow transcript2:37

Then within the body of the slide,

it's good to have an illustrative way to summarize a message you want to convey.

Play video starting at :2:45 and follow transcript2:45

We'll discuss this point specifically in the next video.

But it may be a table, a graph, a chart or an illustration, for instance.

Play video starting at :2:54 and follow transcript2:54

Always add to this illustration some textual explanations.

Your slide should be self-contained, and

you should make sure that all the information the reader needs is on the slides.

Play video starting at :3:7 and follow transcript3:07

Finally, you may decide to have some "take-away" boxes

at the end of some key slides

Play video starting at :3:14 and follow transcript3:14

as a kind of "milestone" within your presentation.

For instance,

a text box emphasizing a key message that your audience should absolutely remember.

At the end of the day, what's difficult is to find the right balance between

self-containedness and conciseness.

In order to be clear, a good stories is neither too brief, nor too long.

So are your slides. And

the content of your slides should be the pieces that structure your story.

Play video starting at :3:44 and follow transcript3:44

I already mentioned that a neat way to make sure your story flows

is to check that all the titles are following each other well.

But you could see it the other way around:

The action title of a specific slide should summarize the slide perfectly.

If you cannot summarize the slide in one sentence,

it's probably that you should split the slide in parts.

One slide, one idea.

# One picture is worth a thousand words

Being able to report your results in a visual way is very important.

First, it will allow you to report a lot of information in a very concise way.

A table or graph may convey a lot more information than text for

the same space on your slide.

Play video starting at ::26 and follow transcript0:26

The same for a visual representation or an image.

A picture is worth a thousand words, plus it's also more convincing.

When seeing your image, if it's well-chosen,

your audience will directly understand what needs to be concluded.

Your audience will see the conclusion by themselves.

It will probably be more convincing than any textual claim you could make.

Play video starting at ::50 and follow transcript0:50

Seeing a decline in sales of a portion in histograms

doesn't need a lot of explanations.

And that's exactly what you want to obtain, this "aha effect".

Play video starting at :1:2 and follow transcript1:02

The best way to learn about data visualization is to be exposed to a lot of

examples and try by yourself to understand what works and what doesn't.

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

That's why in the recital we decided to present a lot of examples to provide you

with a kind of library of data visualization approaches.

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

But one could already keep in mind some principles.

First, if you want your audience to remember your key message,

you want to prune everything that is not absolutely necessary.

That means, that you shouldn't discuss what's not relevant as I said before.

But that also means that you shouldn't show what's irrelevant either.

If a point or a line is not relevant in a graph, remove it.

If a column or a row is not relevant in a table, remove it, and so on.

Play video starting at :1:56 and follow transcript1:56

And by not relevant, I mean that it does not add to your message.

If it's just something interesting but not really useful, it should go away.

Second, Parallelism works.

If you want your audience to compare things, make it comparable.

Place comparable elements at the same level in a table, for instance.

Or if you are using columns to distinguish between different groups in a graph,

use the same color scheme from the beginning to the end, even in the text.

Or always align comparable objects on your slides.

I think you've understood the point.

What I say is that you should always try to be

as consistent as possible in your presentations.

And use balanced structure when you want to emphasize the comparison

between elements.

Third, few people can see in more than two dimensions.

So always try to report your analysis along two dimensions at most.

You will notice that most of the strategic presentations are very often two by two,

the McKinsey matrix, the BCG Matrix, and the likes.

If you cannot avoid working on many dimensions,

use a radar chart for instance,

which is neat way to project several dimensions all at once into a plane.

Play video starting at :3:17 and follow transcript3:17

Finally use images that people will understand directly

by using representations that are standard and as concrete as possible.

Play video starting at :3:28 and follow transcript3:28

Remember the horses and the wild bulls, they were standard and

concrete images, something that people would directly understand.

And so use standard representations as well in your own presentation.

For instance, if you want to report market shares as a function of two factors,

use a bubble chart with the size of the bubble being the market share and

the two axes being the two factors.

This is standard.

Do not reinvent the wheel as the purpose is to make your message

as easy to understand as possible.

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

Many visualizations have been used in the past before you worked on this specific

topic, and use them as much as you can.

That's why we'll try to provide you with

as many examples as possible with the recitals.

Play video starting at :4:19 and follow transcript4:19

But you should continue learning by yourself,

by having your own visualization library in your brain.

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

To conclude, let me repeat again that the objective is to find the right balance

between the quantity of relevant information and

conciseness in order to be clear.

This is very important, but not very easy.

You want your presentation, your work, your story, and

your slides to be as beautiful as they can be.

Play video starting at :4:47 and follow transcript4:47

And always remember that the Temple of Apollo, the Greek god of beauty,

bore the inscription, "meden agan" - "nothing in excess".

It's an aesthetic principle that you could see

as another version of the Occam razor.

If a (a piece) of information is not absolutely needed to convey your message,

you should remove it.

Now it's time for you to practice by yourself again.

We present examples in the recital.

But whatever you could see in this course,

at the end of the day you will have to learn it by yourself.

Data visualization is not a science, it's an art.

Play video starting at :5:25 and follow transcript5:25

Actually it's even more of a craftsmanship.

If you want to improve, you should practice, practice, and practice again.

# Recital M5 - How to present your findings

Hi. My name is Albane Gaubert,

I'm a last year Student at Business School and I'll be your tutor for this recital.

Play video starting at ::7 and follow transcript0:07

In this recital we're going to talk about how to present your findings, and

how to make your point to an audience.

Play video starting at ::15 and follow transcript0:15

First, I'll give you some do's and don'ts,

some basic rules you might want to follow when you present.

Then we'll see some examples, both off slides and of the data visualization.

I hope you enjoy this.

Have a nice MOOC.

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

So, this recital aims to give you a couple of tips

to get your point across during a presentation.

Play video starting at ::36 and follow transcript0:36

The first thing you need to know is that presenting is all

about telling your audience a story they can hear, remember and reproduce.

Play video starting at ::45 and follow transcript0:45

You have to focus on being clear and

making sure they can follow every step of your preference.

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

To do so we are four main things you can do.

The first one is get your audiences attention

by showing them how you are relevant to their problem.

That you perfectly understood why they need you.

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

Best way to do so is to use a pain point.

A pain point is the first thing you're gonna say in a presentation.

It can be a fact, a number, and

it must introduce tension into your demonstration by defining the stake.

Play video starting at :1:22 and follow transcript1:22

Once you have caught your audience's attention, it is all about keeping it.

Play video starting at :1:27 and follow transcript1:27

So, the second must do is connect your ideas throughout the presentation so

that you don't lose any one in the process.

Play video starting at :1:35 and follow transcript1:35

The best way to do that is to structure the presentation as a story so

that each transition makes sense.

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

In case you do lose someone, the best way to get them back is to use action titles.

Play video starting at :1:50 and follow transcript1:50

An action title gives your reader summary of a slide.

In a one shot sentence it serves two purposes.

The first one is to let someone you've lost catch the wagon.

And the second one is help someone quick read through your slides.

Which is very important, especially in case it's students [INAUDIBLE].

Play video starting at :2:11 and follow transcript2:11

Eventually, you have to give actionable recommendations so

that your work matters and has an impact.

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

In order to present effectively there are some rules you might want to

stand by as well.

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

Regarding your presentation, you have to make it clear what the point of each slide

is, and make it easy for the audience to get the most important points.

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

Using the action title we've just talked about coupled with a takeaway box

it's a good way of doing so.

Play video starting at :2:43 and follow transcript2:43

The takeaway box is basically a box where you highlight for the reader what they

should keep in mind after this slide to understand how you come to a next one.

Play video starting at :2:54 and follow transcript2:54

Regarding the contents bear in mind that everything on the slide

will be straight to the point, self containing and relevant.

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

Don't write everything.

This is not a word document but be clear nonetheless.

Play video starting at :3:11 and follow transcript3:11

In order to structure your thoughts you may use bullet points.

But remember that if everything is a bullet point nothing really.

For instance, it is usually a good idea to avoid having bullet

points within bullet points within bullet points.

Play video starting at :3:28 and follow transcript3:28

Last but not least, keep your technique in check.

Remember who you are talking to and wonder whether we can understand this

difficult and precise and fancy with the same technique that you just presented.

And whether they want to, when you do want to explain something a bit technical

Play video starting at :3:48 and follow transcript3:48

or use a method make sure you make it understandable.

Play video starting at :3:54 and follow transcript3:54

With the explained visual method one sentence at most

main stakes without going into details.

If you use a visualization help people read it and

world we live in conclusions for them.

Play video starting at :4:9 and follow transcript4:09

Remember you don't present for yourself but for a person listening to you or

reading your slides another visualization.

You have to remember that your data is only as good as you understand and

communicate it.

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

So make sure you choose the right visualization to have an impact.

Play video starting at :4:32 and follow transcript4:32

The first thing you have to do, like of a presentation is to identify the story

to define your data so I can pick the right visualization for it.

Play video starting at :4:45 and follow transcript4:45

It is important to identify, and understand the story you are trying to

tell, and the relationship you're looking to show.

Play video starting at :4:53 and follow transcript4:53

Knowing this information will help you select the proper visualization to best

to best deliver your message.

Play video starting at :5: and follow transcript5:00

Data visualization also has a couple of do's and don'ts and it's best to follow.

Play video starting at :5:7 and follow transcript5:07

For instance,

Play video starting at :5:9 and follow transcript5:09

use only one color to represent each category otherwise gets messy.

And difficult to understand.

Order data sets using logical view key may be from biggest to a smallest or

be very round or in terms of categories but you have to have a hierarchy.

Play video starting at :5:29 and follow transcript5:29

You can use call outs to highlight important or

interesting informations Visualize data in a way that is easy for

the reader to compare values and that comes with VRK.

Play video starting at :5:42 and follow transcript5:42

Use icons to enhance comprehension and reduce unnecessary labeling.

Play video starting at :5:48 and follow transcript5:48

Can use high contrast color combinations.

Play video starting at :5:52 and follow transcript5:52

Do not use on the other side high contrast color combinations, so

there's such as red or green or blue and yellow.

Play video starting at :6:1 and follow transcript6:01

Is the eye.

Don't use 3D charts this perception of visualization

Play video starting at :6:8 and follow transcript6:08

but if you don't get to see that as well and it's distracting for the reader.

Play video starting at :6:13 and follow transcript6:13

Don't add chart check even.

And this is how illustration drop shadows or

ornamentation distract from the data as well.

Play video starting at :6:23 and follow transcript6:23

Don't show more than six columns single layout and

once we get the same point if it's too messy or

too different shaded afterwards it gets confusing for the reader.

Play video starting at :6:33 and follow transcript6:33

Don't use distracting fonts or elements the same points till which is bold,

italic, or underline text.

Play video starting at :6:45 and follow transcript6:45

So for starters you look at this slide you see we have a pinpoint as

the very first thing we tell the audience.

Play video starting at :6:53 and follow transcript6:53

Basically we define the stage for inefficience of a media strategy and

this is why we're here.

We want to fix this.

Play video starting at :7:3 and follow transcript7:03

On this next slide consider who's thought of an action title

Play video starting at :7:7 and follow transcript7:07

that defines what we're going to talk about AVI scores.

Play video starting at :7:12 and follow transcript7:12

We ultimately find the method for AVI.

That's what AVI scores and how we computed them.

Can see here that we keep it very to the top and simple.

Only one line and one line also an hypothesis.

Play video starting at :7:30 and follow transcript7:30

Here we also have a takeaway box.

Finding what we've learned at the end of the slide.

Play video starting at :7:37 and follow transcript7:37

This slide is also the same structure.

First an action title then the method how it will get there,

how did you do, what's these graph mean?

Play video starting at :7:52 and follow transcript7:52

Then as we have a graph, we give it a title, what is it, what is the timeframe.

Play video starting at :7:59 and follow transcript7:59

A legend which has to be consistent so for

instance if you two things that are equivalent but

[INAUDIBLE] different you use the same colors and use the same codes.

Then you try to reduce the number of different colors you use,

you give the main point.

Here, for instance, it's 100 AVI because if it's under 100 it

means it's not efficient, if it is over it, it means it is efficient.

Play video starting at :8:26 and follow transcript8:26

You give findings for the visualization you gave.

And a conclusion to a slide of all.

You can also find great examples of data visualizations

in the press or in other places, The Economist for instance is a great source.

You can see here that we have a simple title, simple legend,

few colors had the axis are defined and that we are the source.

Play video starting at :8:55 and follow transcript8:55

We are also given the R two and the countries are ordered,

the continents are ordered in alphabetical order.

Play video starting at :9:6 and follow transcript9:06

This graph is also really good because it's very to the point.

We can see the home index per state from 1971 to 2013 and

have the evolution from that.

Play video starting at :9:19 and follow transcript9:19

This is about the same thing like we have a title, we have the time frame,

we have a scale, ticks indicate volume in migration, of migration in millions.

Play video starting at :9:33 and follow transcript9:33

And we have the graph itself, that indicates all that.

And that's ordered as well.

Play video starting at :9:43 and follow transcript9:43

This graph is interesting as well.

It is very to a point.

We see the airline routes across the world basically.

It is interesting to note that the airlines have been ordered

in a way that it's not immediately understandable.

So they've actually explained why it's ordered that way.

Play video starting at :10:5 and follow transcript10:05

We have the different colors, not too many of them so

that we can still read the graph basically.

Play video starting at :10:13 and follow transcript10:13

And this one is basically one of the best one you can get.

You have a title, then the axises are defined, it show selected cities and

the average costs of buy a car and renting a family car in each city.

Here are the cities.

Here are the costs.

They split the costs in two colors, for purchase price and the running costs.

They give you a scale that starts at 0, which is really important for

you to actually get the big picture.

And the evolution, so that you know the long dynamic thing.

Play video starting at :10:50 and follow transcript10:50

And we have a source here as well so

that you can actually place some outside data source.

Play video starting at :10:57 and follow transcript10:57

So, basically that was all.

I hope it was clear and that you understood most of what I said and

that you will be able to reuse it sometime.

Play video starting at :11:5 and follow transcript11:05

Good luck.

# Presentation Tips

[Presentation tips](https://d3c33hcgiwev3.cloudfront.net/_661e8cedb5386691a82af9db2c9f0c65_Presentation-tips.pptx?Expires=1703721600&Signature=LGOWnT3y6fUvfmPNMzCOcoSwgzkyP5PBLzvvCcBs0qirC-4yyY5Mc78JEqgcOafHpBlb~H5TfaKGyvLhHPW8FGlzP8mFS636tQ5wcU1pl3plZ8qVu~2foHyw8U4i8hb52igLW-3AcZaRvSIzq4sLhXcD89Z1OPU53gYPQ-aV88M_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[PPTX File](https://d3c33hcgiwev3.cloudfront.net/_661e8cedb5386691a82af9db2c9f0c65_Presentation-tips.pptx?Expires=1703721600&Signature=LGOWnT3y6fUvfmPNMzCOcoSwgzkyP5PBLzvvCcBs0qirC-4yyY5Mc78JEqgcOafHpBlb~H5TfaKGyvLhHPW8FGlzP8mFS636tQ5wcU1pl3plZ8qVu~2foHyw8U4i8hb52igLW-3AcZaRvSIzq4sLhXcD89Z1OPU53gYPQ-aV88M_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

# Wrap-up: reporting your results

In this module,

we emphasized why storytelling is important in a business analytics context.

Play video starting at ::16 and follow transcript0:16

First, you want to start with a hook up,

emphasizing how important the issue you raise is.

Play video starting at ::23 and follow transcript0:23

Then your presentation needs to flow and come naturally, and

logically to your conclusions.

We underline that one key success factor of implementing your recommendations

will be that you're getting your audience to have a shared representation

of the reality.

Play video starting at ::42 and follow transcript0:42

And this will ensure that everybody comes to the same conclusions

by the end of your presentation.

If they don't come to the same conclusions,

they won't act together to implement you recommendations.

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

So this point is crucial for

ensuring the success of your influence as a business analytics expert.

Finally, we also discussed how to structure those slides.

All your slides should be as visual as possible, while also self-contained.

And I said that the best way to communicate well is to use images and

representations that speak by themselves.

As from now on,

you should practice and try to identify those images that work well.

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

And that's something you can train yourself for

by being exposed to a lot of different examples.

Now for the last exercise of this mode,

we would like you to practice a qualification issue on actual data.

Play video starting at :1:36 and follow transcript1:36

And we'll kill two birds with one stone

by already preparing the capstone project of this specialization.

We prepared a list of data sets that you can work on.

We'll find the documentation describing those datasets, and

simple positions of issues that could be addressed.

Note that, if you find more interesting data set equations,

you should be free to use it, as long as the data is accessible to everybody.

Play video starting at :2:1 and follow transcript2:01

The objective is that any team may decide to replicate the results of another team.

Play video starting at :2:7 and follow transcript2:07

That's why all the data sets that we list are open sourced.

Play video starting at :2:12 and follow transcript2:12

So, for the end of the course, we would like you to prepare

the first slide of your capstone project presentation.

Play video starting at :2:20 and follow transcript2:20

It may be a "pain point" to solve, or an opportunity to take, but

this one-page needs to be relevant, visual, self-contained, and clear.

In short, everything we emphasized in this module.

You will then be assessed by your peers

Based on four criteria. Each time assessed on a scale of one to five,

one being very bad and five being very good.

1/ how relevant is the issue rate for a business or an organization?

Play video starting at :2:48 and follow transcript2:48

2/ how well is the issue presented,

in a visual and quantitative way?

Play video starting at :2:54 and follow transcript2:54

3/ Is it the right level of information?

Not too much, neither too little?

4/ is the slide well-structured?

Do we have a hierarchy of information that allows us to read the slide

at different levels of depth?

Play video starting at :3:8 and follow transcript3:08

You've made your first steps as a business analytics consultant.

We've discussed some tools and examples,

but we just unveiled the tip of the iceberg.

In other MOOCs in this training,

you will have the opportunity to see methodologies in more detail,

such as during the Marketing Analytics course,

or cover a wider range of examples from various industries,

such as, during the case studies with Accenture training.

Play video starting at :3:35 and follow transcript3:35

Also, do not hesitate to share your ideas or

concerns with the other participants of this MOOC on the forum.

Play video starting at :3:43 and follow transcript3:43

Those exchanges are likely to foster interesting dynamics

leading to the development of your own Business Analytics skills.

Play video starting at :3:53 and follow transcript3:53

Now to conclude

I'd like to say that I was very happy to do this training on

Strategic Business Analytics with you and I wish you the best of luck for

your Strategic Business Analytics career.

# Datasets for Peer Review Assignment

1. **Monitoring of CO2 emissions from passenger cars – Regulation 443/2009**

European Environment Agency

The Regulation (EC) No 443/2009 requires Member States to record information for each new passenger car registered in its territory. Every year, each Member State shall submit to the Commission all the information related to their new registrations. In particular, the following details are required for each new passenger car registered: manufacturer name, type approval number, type, variant, version, make and commercial name, specific emissions of CO2, mass of the vehicle, wheel base, track width, engine capacity, fuel type and fuel mode. Additional information, such as engine power, were also submitted.

<http://www.eea.europa.eu/data-and-maps/data/co2-cars-emission-8>

TRANSPORT – CAR – ENVIRONMENT – REGULATION

**2. Speed Dating**

Department of Statistics – Columbia University

Speed dating data with over 8,000 observations of matches and non-matches, with answers to survey questions about how people rate themselves and how they rate others on several dimensions. This is a large and rich dataset which might take you some time to fully understand. It should be fun to play with.

<http://www.stat.columbia.edu/~gelman/arm/examples/speed.dating/>

FUN – DATING – HUMAN

**3. Bike Sharing**

Laboratory of Artificial Intelligence and Decision Support (LIAAD), University of Porto

Apart from interesting real world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for the research. Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns bike sharing system into a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of important events in the city could be detected via monitoring these data. The dataset contains 17389 instances and 16 attributes.

**Erratum**: in dataset description - season : 1=winter, 2=spring, 3=summer, 4=fall

[https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset#](https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset)

BIKE – TRANSPORT – LOGISTIC

**4. Loans**

Lending Club Corporation

These files contain complete loan data for all loans issued through the time period stated, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. The file containing loan data through the "present" contains complete loan data for all loans issued through the previous completed calendar quarter. Another dataset contains the list and details of all loan applications that did not meet Lending Club's credit underwriting policy. You have to Sign in to download the full version of the files.

<https://www.lendingclub.com/info/download-data.action>

FINANCE – BANK – BUSINESS

**5. OpenFlights – Airport, Airline & Route Database**

OpenFlights

The OpenFlights Airports Database contains 6977 airports spanning the globe.

The OpenFlights Airlines Database contains 5888 airlines.

The OpenFlights/Airline Route Mapper Route Database contains 59036 routes between 3209 airports on 531 airlines spanning the globe.

<http://openflights.org/data.html>

TRANSPORT – AIRLINE – AIRPORT – ROUTE

**6. The Insurance Company Benchmark – KDD Cup**

Information and Computer Science, University of California, Irvine

This data set contains information on customers of an insurance company. The data consists of 86 variables and includes product usage data and socio-demographic data derived from zip area codes. The data was collected to answer the following question: Can you predict who would be interested in buying a caravan insurance policy and give an explanation why?

<http://kdd.ics.uci.edu/databases/tic/tic.html>

INSURANCE – BUSINESS – MARKETING

**7. Mailing campaign for NPO – KDD Cup 1998**

Information and Computer Science, University of California, Irvine

The dataset consists in a regression problem where the goal is to estimate the return from a direct mailing in order to maximize donation profits.

<http://kdd.ics.uci.edu/databases/kddcup98/kddcup98.html>

NON-PROFIT – CAMPAIGN – MARKETING

**8. Customer relationship prediction – KDD Cup 2009**

KDD Cup / Orange

This dataset offers the opportunity to work on large marketing databases from the French Telecom company Orange to predict the propensity of customers to switch provider (churn), buy new products or services (appetency), or buy upgrades or add-ons proposed to them to make the sale more profitable (up-selling). Both training and test sets contain 50,000 examples. For the large dataset, the first 14,740 variables are numerical and the last 260 are categorical. For the small dataset, the first 190 variables are numerical and the last 40 are categorical.

<http://kdd.org/kdd-cup/view/kdd-cup-2009/Data>

CHURN – CUSTOMER RELATIONSHIP – MARKETING

**9. Fuel prices**

ETALAB, data.gouv.fr

The dataset consists in daily prices for gas stations in France from 2007 to 2014. It contains information such as the address, geographical information, working hours, prices, services provided and permanent or temporary closure if it is the case. It also contains historical information to allow comparisons.

<https://www.data.gouv.fr/en/datasets/prix-des-carburants-en-france/>

FUEL – GEOGRAPHICAL – ENERGY

**10. Medical expense refunds (Medicam)**

ETALAB, data.gouv.fr

The Medic'AM dataset reports the medical expenses refunds by the French health insurance. For each medicament, the dataset provides its name, its category, the refunded basis, the number of refunded medicaments, the refunded amount and the prescribers basis. The dataset contains data from 2008 to 2013.

<https://www.data.gouv.fr/en/datasets/medicaments-rembourses-par-lassurance-maladie/>

HEALTH – INSURANCE – FRAUD

**11. Establishment Specific Injury & Illness Data (OSHA Data Initiative)**

United State Departement of Labor

The Occupational Safety and Health Administration (OSHA) collected work-related injury and illness data from employers within specific industry and employment size specifications from 1996 through 2011. The data provided is used by OSHA to calculate establishment specific injury and illness incidence rates. This searchable database contains a table with the name, address, industry, and associated Total Case Rate (TCR), Days Away, Restricted & Transfer (DART) case rate, and the Days Away From Work (DAFWII) case rate for the establishments that provided OSHA with valid data.

<https://www.osha.gov/pls/odi/establishment_search.html>

INJURY – MEDICAL – LABOUR

Disclaimer: those datasets are provided by third parties. They can be made unavailable by those parties without notice. ESSEC is not liable in any case for any problem occurred because of any type of usage of those datasets.

# Peer-graded Assignment: Peer Review

DeadlineJan 21, 11:59 PM PST

i

It looks like this is your first peer-graded assignment. [Learn more](https://learner.coursera.help/hc/articles/208279926-Submit-peer-reviewed-assignments)

**Ready for the assignment?**

You will find instructions below to submit.

1. [**Instructions**](https://www.coursera.org/learn/strategic-business-analytics/peer/V6J6s/peer-review)
2. [**My submission**](https://www.coursera.org/learn/strategic-business-analytics/peer/V6J6s/peer-review/submit)
3. [**Discussions**](https://www.coursera.org/learn/strategic-business-analytics/peer/V6J6s/peer-review/discussions)

This deliverable objective is to be a first draft of one of the first slides of your presentation for the capstone project. This one-pager needs to describe what your project will be addressing. It may be a « pain point » to solve, or an opportunity to take. But it needs to be relevant, visual, self-contained, and clear.

### Grading Criteria Overview

Review 4 submissions by your peers in order to pass this assignment

### Instructions

For this last exercise of this course, we would like you to practice a qualification issue on actual data. And we’ll kill two birds with one stone, by already preparing the capstone project of this specialisation.

We’ve prepared a list of datasets (<http://bit.ly/1F9YKSD>) that you can work on. You’ll find a documentation describing those datasets and some propositions of issues that could be addressed. Note that if you find more interesting datasets or questions, you should feel free to use it as long as the data is accessible to everybody. The objective is that any other participant may decide to replicate the results of another participant. That’s why all the datasets that we list are open source...

So, for the end of this course, we would like you to already prepare the first slide of your capstone project presentation. It may be a « pain point » to solve, or an opportunity to take. This is an issue qualification problem. Not a question where there is a problem to solve or a correct answer to find... This one page needs to be relevant, visual, self-contained, and clear. In short, everything we emphasized in the module 5.

The document needs to be one page long maximum with a font size of 12 minimum. Please provide a pdf as this will ensure that there will be no versioning problem. Make sure that all the content of your slide is readable and well presented.

In order to begin the data investigation process, you can first start replicating the different analyses discussed during the course on your own data set: can you find groups of similar or dissimilar observations? Can you find relationships between causes and consequences? Always try to be produce some visuals as we did during the videos... You'll probably notice patterns that are relevant or surprising. This may lead to relevant opportunities to take or problems to solve. Also try to select a data set related to a business you are familiar with or that you can discuss with your friends or colleagues. Do not hesitate to interview people from this specific sector: they may be able to tell you what could be relevant to investigate... In any case, do not hesitate to ask questions on the forum and to discuss your ideas or insights with the other participants.

### Review Criteria

You will then be assessed by your peers based on 4 criteria. Each time assessed on a scale of 1 to 5, 1 being very bad, and 5 being very good:

1. Relevance: How relevant is the issue raised for a business or an organization?

2. Presentation: How well is the issue presented, in a visual and quantitative way?

3. Level of information: Is it the right level of information? Nor too much, neither too little?

4. Structure: Is the slide well structured? Do we have a hierarchy of information that allows to read the slide at different level of depth?

When you will evaluate your peers, you will be asked to motivate your assessment by indicating:

1. 3 "attention points" that you think should have been improved with regard to this criterion

2. 3 "strengths" of the deliverable with regard to this criterion

You need to present the situation emphasizing the pain point or opportunity in a data-based and visual way. You don't need to produce any advanced analysis or conclusion yet, but the presentation of the problem needs to be factual, relevant and visual. For instance, in the case of the HR Analytics example, a figure reporting the defection would be relevant. Another example, in the predictive maintenance case, would be to report the number of broken elements and to emphasize what would be the expected benefit of a prediction analysis. It's the relevance of the issue that needs to be emphasized, not the analyses or the conclusions yet...

This is an issue qualification problem. Not a question where there is a problem to solve or a correct answer to find

<https://rpubs.com/reemsoliman/328754>

<https://github.com/nicolasfguillaume/Strategic-Business-Analytics-with-R/blob/master/module4.md>