**DIGITAL ANALYTICS AND REGRESSION**

### About this course

**Learning Objectives**

In this course you will learn:

* The entire business analytics cycle including
  1. Defining a business objective
  2. Creating a data analytics solution
  3. Articulating a recommendation

As you work through the case study and exercises, start thinking about how the 'analytics cycle' can be applied in different situations. This approach can be used in situations where historical data has been tracked and stored and, where that data is available for extraction e.g., app downloads, marathon running times, or product sales.

Dans ce cours, vous apprendrez :

* Le cycle complet de l'analyse commerciale, y compris
  1. Définition d'un objectif commercial
  2. Créer une solution d'analyse des données
  3. Formulation d'une recommandation

Au fil de l'étude de cas et des exercices, commencez à réfléchir à la manière dont le "cycle analytique" peut être appliqué dans différentes situations. Cette approche peut être utilisée dans des situations où des données historiques ont été suivies et stockées et où ces données sont disponibles pour l'extraction, par exemple pour le téléchargement d'applications, les temps de course du marathon ou la vente de produits.

### Module 1: Case Study 'CEO vs. CMO'

**Learning Objectives**

**Lesson 1 – A Case Study Approach to Analytics**

* Understand the business context
* Formulate the business objective
* State the hypothesis
* Assess available data
* Assign data for use

Leçon 1 - Une approche analytique par étude de cas

* Comprendre le contexte de l'entreprise
* Formuler l'objectif de l'entreprise
* Énoncer l'hypothèse
* Évaluer les données disponibles
* Attribuer les données à utiliser

## Case Assignment: 'CEO' vs. 'CMO'

You are a newly-hired data science analyst at a small tech startup. The company you have just joined has created a new cloud based technology that allows users to access R and other programming tools through any web browser. This new technology has 5 main advantages:

Vous venez d'être embauché comme analyste en sciences des données dans une petite start-up technologique. La société que vous venez de rejoindre a créé une nouvelle technologie basée sur le cloud qui permet aux utilisateurs d'accéder à R et à d'autres outils de programmation par le biais de n'importe quel navigateur web. Cette nouvelle technologie présente 5 avantages principaux :

1. Cost savings - reduced hardware upgrade costs
2. Increased data security - encryption, backup and disaster recovery
3. The technology allows for greater functionality, efficiency and innovation in data science projects
4. More flexibility - access to data and data science projects anytime, anywhere and from any device with an internet connection. This technology will allow data scientists all over the world to run sophisticated statistical models on the go.
5. Even more flexibility - no more installations, version control issues and, no need to choose a CRAN mirror
6. Réduction des coûts - réduction des coûts de mise à niveau du matériel
7. Sécurité accrue des données - cryptage, sauvegarde et récupération après sinistre
8. Cette technologie permet une plus grande fonctionnalité, efficacité et innovation dans les projets de science des données
9. Plus de flexibilité - accès aux données et aux projets scientifiques à tout moment, en tout lieu et depuis n'importe quel appareil disposant d'une connexion Internet. Cette technologie permettra aux spécialistes des données du monde entier d'exécuter des modèles statistiques sophistiqués lors de leurs déplacements.
10. Encore plus de flexibilité - plus d'installations, plus de problèmes de contrôle de version et, pas besoin de choisir un miroir CRAN

*The programming language R requires users to choose a cran mirror before executing certain commands.*A mirror is a distribution site for software. *In addition to a main download site, a mirror has a identical copies of the software (libraries, source code or documentation) to be downloaded. This allows a user to choose a download site that is geographically near to them, improving the bandwidth or latency, it balances the load across many servers and, it provides redundancy so that the software is available even if one or more of the mirror sites is unavailable. A main download site is not required, it is possible to distribute software via a network of mirror sites.*

Le langage de programmation R exige des utilisateurs qu'ils choisissent un miroir cranté avant d'exécuter certaines commandes. Un miroir est un site de distribution de logiciels. En plus d'un site principal de téléchargement, un miroir dispose d'une copie identique des logiciels (bibliothèques, code source ou documentation) à télécharger. Cela permet à l'utilisateur de choisir un site de téléchargement géographiquement proche de lui, ce qui améliore la bande passante ou la latence, équilibre la charge sur de nombreux serveurs et assure une redondance de sorte que le logiciel est disponible même si un ou plusieurs des sites miroirs sont indisponibles. Un site principal de téléchargement n'est pas nécessaire, il est possible de distribuer le logiciel via un réseau de sites miroirs.

In your first team meeting, and one week before the scheduled product launch, the Chief Executive Officer (CEO) and the Chief Marketing Officer (CMO) have a heated argument about what to call the product. The implications of this decision are significant, everything from their social media accounts, to the product website; their whole identity and positioning will be directly linked to this name.

Lors de votre première réunion d'équipe, et une semaine avant le lancement prévu du produit, le directeur général (PDG) et le directeur du marketing (CMO) ont une discussion animée sur le nom à donner au produit. Les implications de cette décision sont importantes, de leur compte sur les médias sociaux au site web du produit ; toute leur identité et leur positionnement seront directement liés à ce nom.

After much deliberation, the two executives have narrowed it down to just 2 options:

Après de longues délibérations, les deux dirigeants ont réduit le choix à deux options seulement :

1. **Option A: Analytics Workbench**
2. **Option B: Data Scientist Workbench**
3. Option A : Analytics Workbench
4. Option B : Data Scientist Workbench

The CEO and the CMO both agree that the target audience for the product consists of data scientists and aspiring data scientists. These people do work in machine learning, data mining, modeling and visualization. They use data to answer questions and solve problems.

Le PDG et l'OCM s'accordent à dire que le public cible du produit est constitué de scientifiques de données et de scientifiques de données en herbe. Ces personnes travaillent dans l'apprentissage machine, l'exploration de données, la modélisation et la visualisation. Ils utilisent les données pour répondre à des questions et résoudre des problèmes.

The CEO has a vision to position the company as a global leader in data science and with a limited marketing budget, he knows that good Search Engine Optimizaton (SEO) and online advertising, Search Engine Marketing (SEM) will be critical for trial and adoption of the product. He believes that the two names are actually interchangeable and they will attract similar people; after all, someone searching for ‘Analytics’ would surely be interested in the big data wave that’s sweeping the world. He has done a quick search on a popular search engine tool, Google Trends, and found that ‘Analytics’ is, by far, a more popular search term than ‘Data Science’ or ‘Data Scientist’. The CEO is making a data driven decision to go with Option A, 'Analytics Workbench'.

Le PDG a pour vision de positionner l'entreprise comme un leader mondial dans le domaine de la science des données et, avec un budget marketing limité, il sait qu'une bonne optimisation pour les moteurs de recherche (SEO) et la publicité en ligne, le marketing pour les moteurs de recherche (SEM) seront essentiels pour l'essai et l'adoption du produit. Il pense que les deux noms sont en fait interchangeables et qu'ils attireront des personnes similaires ; après tout, quelqu'un qui recherche des "Analytiques" serait sûrement intéressé par la grande vague de données qui balaie le monde. Il a fait une recherche rapide sur un outil de recherche populaire, Google Trends, et a trouvé que "Analytics" est, de loin, un terme de recherche plus populaire que "Data Science" ou "Data Scientist". Le PDG prend une décision basée sur les données pour choisir l'option A, "Analytics Workbench".

The CMO, who has some experience in SEM from past work at another tech company, has a gut feeling that Option B: Data Scientist Workbench, will bring the right target group to the new product. He knows that, their small marketing budget won’t go very far if they cast the net too wide with a very popular search term! They would have to make higher bids for keywords that may have a lot of competing offers and, the bigger audience would generate more clicks including clicks from people that are not looking for data science related content. They would be billed for every click! Ideally their product website should be the first item to pop up when people search for data science related content and their online ads should be optimized to deliver the right message to the right audience for the best quality traffic to their website. He feels strongly that the product name should align more closely with their target audience.

Le directeur général, qui a une certaine expérience du SEM grâce à son travail passé dans une autre entreprise technologique, a le sentiment que l'option B, "Data Scientist Workbench", permettra d'atteindre le bon groupe cible pour le nouveau produit. Il sait que son petit budget de marketing n'ira pas très loin s'il jette le filet trop large avec un terme de recherche très populaire ! Ils devraient faire des offres plus élevées pour des mots-clés qui peuvent avoir beaucoup d'offres concurrentes et, le public plus large générerait plus de clics, y compris des clics de personnes qui ne recherchent pas de contenu lié aux sciences des données. Ils seraient facturés pour chaque clic ! Idéalement, le site web de leur produit devrait être le premier élément à apparaître lorsque les gens recherchent un contenu lié aux sciences des données et leurs annonces en ligne devraient être optimisées pour délivrer le bon message au bon public pour un trafic de meilleure qualité vers leur site web. Il est convaincu que le nom du produit devrait être plus proche de son public cible.

The seasoned CMO would have loved to do some AB testing or to conduct a focus group or survey to test the quality of traffic to both names but there just wasn't any time or budget for that. To make matters worse, the company selling the website domains with those two names will double their price in 4 hours and wants an answer in 8 hours, otherwise they will sell both domains to a competitor. The CMO trusts his gut that ‘Option B - Data Scientist Workbench’ is the way to go but he can’t argue with the data, the more popular search term is ‘analytics’.

L'OCM chevronné aurait aimé faire des tests AB ou mener un groupe de discussion ou une enquête pour tester la qualité du trafic vers les deux noms, mais il n'y avait tout simplement pas de temps ni de budget pour cela. Pour aggraver les choses, la société qui vend les domaines de sites web avec ces deux noms doublera leur prix en 4 heures et veut une réponse en 8 heures, sinon elle vendra les deux domaines à un concurrent. L'OCM se fie à son instinct pour dire que l'"option B - Data Scientist Workbench" est la meilleure solution, mais il ne peut pas contester les données, le terme de recherche le plus populaire étant "analyse".

**Given your background and experience, the two executives request that you provide some insight to help them make this decision.**

Compte tenu de vos antécédents et de votre expérience, les deux dirigeants vous demandent de leur fournir des informations pour les aider à prendre cette décision.

## Develop the Business Objective

### Developing the business objective for an analysis is a process of asking and answering a series of questions. Some of the questions we’ll need to ask ourselves in determining the business objectives are:

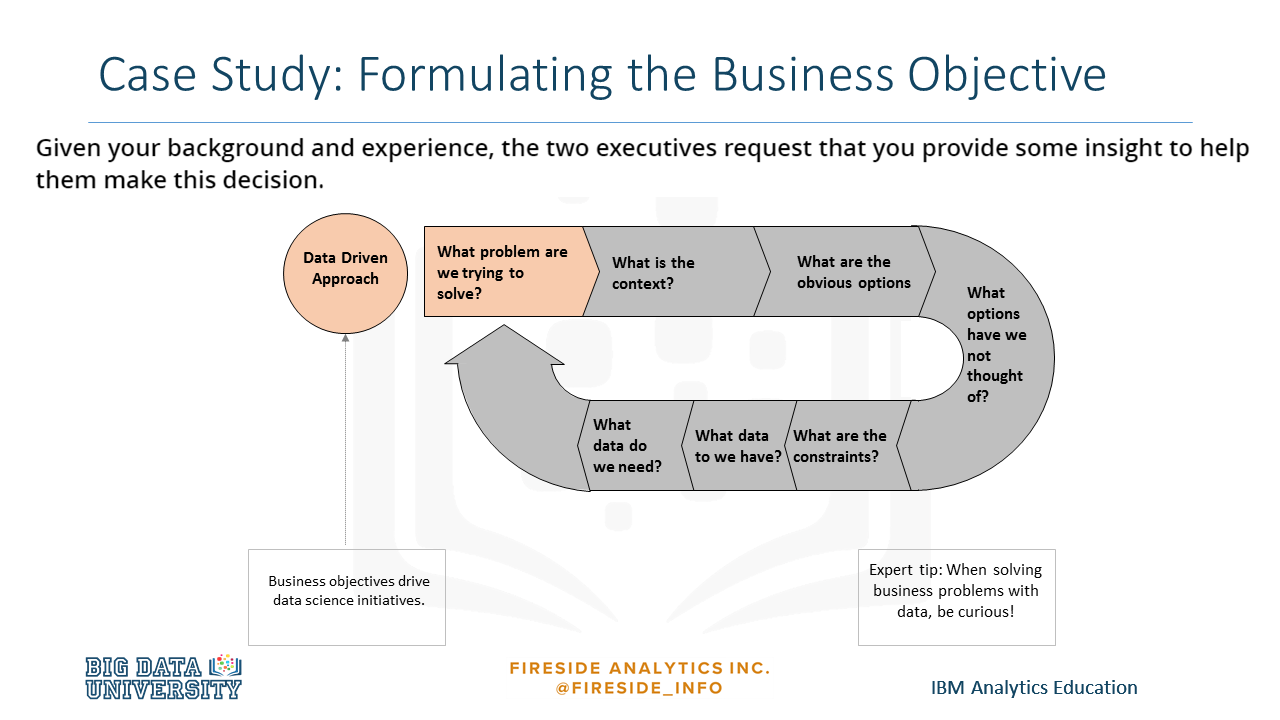
• What is the goal of the solution?  
• Why do we need to do this?  
• What data is available?  
• What constraints exist?  
• What is an acceptable solution?  
• What is success in this context?  
• How do we measure success?

L'élaboration de l'objectif commercial d'une analyse est un processus consistant à poser et à répondre à une série de questions. Voici quelques-unes des questions que nous devrons nous poser pour déterminer les objectifs commerciaux :

* Quel est le but de la solution ?
* Pourquoi devons-nous le faire ?
* Quelles sont les données disponibles ?
* Quelles sont les contraintes existantes ?
* Quelle est une solution acceptable ?
* Qu'est-ce que le succès dans ce contexte ?
* Comment mesurer le succès ?

### Developing the business objective for an analysis may follow these useful steps:

### L'élaboration de l'objectif commercial pour une analyse peut suivre ces étapes utiles :



**(1)  What is our objective? E.g., What is the business problem we are trying to solve? What decision needs to be made?**

(1) Quel est notre objectif ? Par exemple, quel est le problème commercial que nous essayons de résoudre ? Quelle décision doit être prise ?

In this situation, think of the business problem as a statement to prove or disprove or an hypothesis to accept or reject:

Dans cette situation, considérez le problème commercial comme une déclaration à prouver ou à réfuter ou une hypothèse à accepter ou à rejeter :

**e.g., The product name should be Option A, ‘Analytics Workbench’ because Google Trends data showed that there are more searches for ‘Analytics’ than for ‘Data Science’.**

Par exemple, le nom du produit devrait être l'option A, "Analytics Workbench", car les données de Google Trends ont montré qu'il y a plus de recherches pour "Analytics" que pour "Data Science".

Is this true? If so, we may go back to our team and advise them that the data does indeed support the CEO’s decision. The heart of the question however is asking which search term is ‘better’. This would take more work on our part to answer. We can proceed with our analysis by accepting or rejecting the CEO’s decision and then answering the heart of the question, ‘which term is better?’.

Est-ce vrai ? Si c'est le cas, nous pouvons retourner voir notre équipe et l'informer que les données soutiennent effectivement la décision du PDG. Le cœur de la question est toutefois de savoir quel terme de recherche est "meilleur". Il nous faudrait davantage de travail pour répondre à cette question. Nous pouvons poursuivre notre analyse en acceptant ou en rejetant la décision du PDG et en répondant ensuite au cœur de la question, "quel terme est le meilleur".

**(2)  What is the context? E.g., are there any noteworthy people dynamics? Who is your champion for the analysis? Whose problem are you solving? Who stands to gain or be exposed from the outcome of your analysis? Does anybody care?**

(2) Quel est le contexte ? Par exemple, y a-t-il une dynamique de personnes remarquables ? Qui est votre champion pour l'analyse ? Quel est le problème que vous résolvez ? Qui devrait bénéficier ou être exposé aux résultats de votre analyse ? Est-ce que quelqu'un s'en soucie ?

**Context and assumptions:**

Contexte et hypothèses :

When embarking on a new analytics project with a new team, take in as much information from your team as possible. This can be a formalized process of interviewing different team members or in casual conversations.

Lorsque vous vous lancez dans un nouveau projet d'analyse avec une nouvelle équipe, recueillez autant d'informations que possible auprès de votre équipe. Il peut s'agir d'un processus formalisé consistant à interroger différents membres de l'équipe ou de conversations informelles.

**\*\*\*Expert Tip\*\*\***

\*\*\*Conseil d'expert\*\*\*

Information from team members may come with unconscious biases, accept nothing as fact that you cannot reasonably independently verify. While being careful not to do 'rework', it is important to verify any ideas that are the basis of your analysis. Sometimes assumptions presented as facts at the very beginning of your process can send you down an analysis path that is incorrect.

Les informations fournies par les membres de l'équipe peuvent être entachées de préjugés inconscients, n'acceptant rien comme un fait que vous ne pouvez raisonnablement pas vérifier de manière indépendante. Tout en prenant soin de ne pas faire de "retouches", il est important de vérifier les idées qui sont à la base de votre analyse. Parfois, des hypothèses présentées comme des faits au tout début de votre processus peuvent vous envoyer sur une voie d'analyse qui est incorrecte.

**When solving business problems with data, be curious!**

Lorsque vous résolvez des problèmes commerciaux avec des données, soyez curieux !

In this case, the CEO has given us a statement, “‘Analytics’ is the better performing search term,” however, if we are to base our entire analysis on this idea, we should verify this statement independently as part of your analysis.

Dans ce cas, le PDG nous a donné une déclaration, "'Analytique' est le terme de recherche le plus performant," cependant, si nous devons baser toute notre analyse sur cette idée, nous devrions vérifier cette déclaration indépendamment dans le cadre de votre analyse.

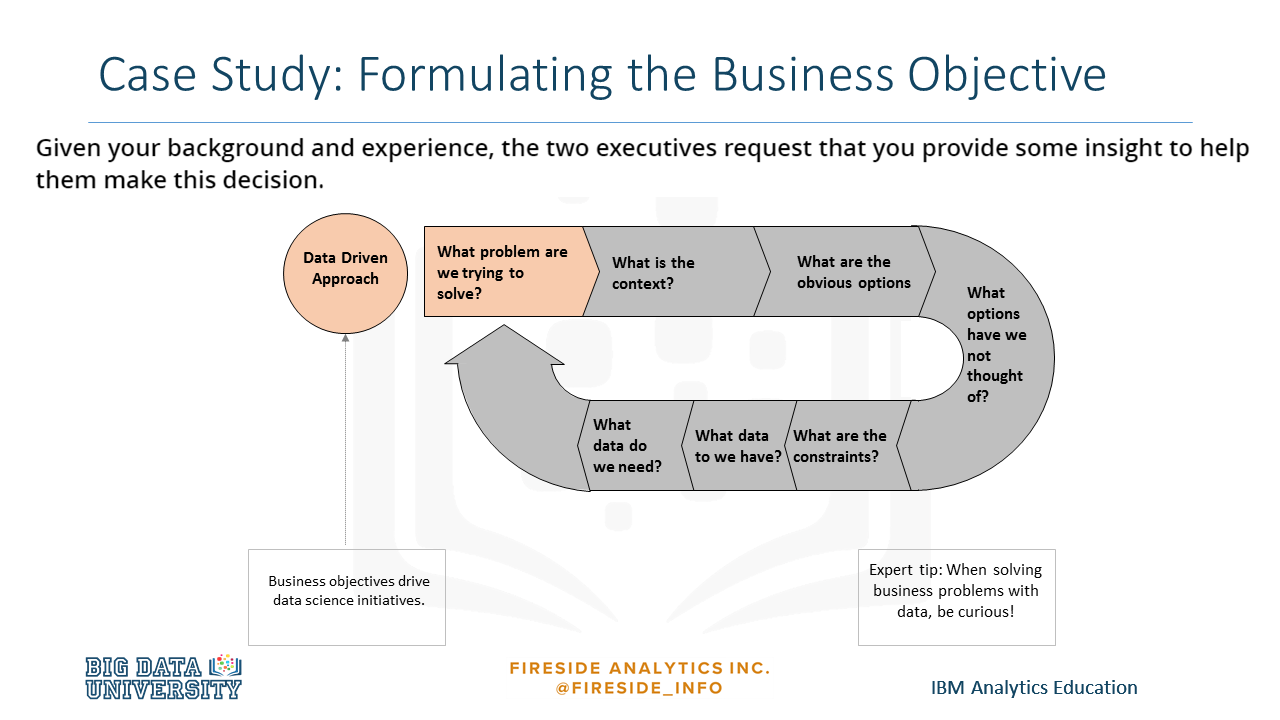
## A Data Driven Approach

**Developing the business objective for an analysis may follow these useful steps:**

continued..

L'élaboration de l'objectif commercial pour une analyse peut suivre ces étapes utiles :

suite...

****

**(3)  What are the obvious options?**

Option A: Analytics Workbench

Option B: Data Scientist Workbench

(3) Quelles sont les options évidentes ?

Option A : Analytics Workbench

Option B : Data Scientist Workbench

**(4)  What have we not yet considered?**

* Which search term more accurately reflects the correct target audience?
* What other data comes in a Google Trends file?
* Are there any other name options that are better than both initial options?

(4) Qu'avons-nous encore négligé ?

* Quel est le terme de recherche qui reflète le mieux le bon public cible ?
* Quelles autres données figurent dans un fichier Google Trends ?
* Y a-t-il d'autres options de nom qui sont meilleures que les deux options initiales ?

**(5)  What are the constraints?**

In any situation, there are a number of constraints that need to be managed and taken into account for a meaningful analysis.

(5) Quelles sont les contraintes ?

Dans toute situation, il existe un certain nombre de contraintes qui doivent être gérées et prises en compte pour une analyse pertinente.

* Budget constraints

1. SEM can become expensive, and the related efforts of gathering and processing data can be very expensive. Some advertising agencies are paid as a percentage of advertising spend, which does not incentivize them to keep advertising costs down e.g., even if conversions are high, the cost per conversion must be minimized.
2. Many organizations may lack the expertise to conduct extensive surveys to gather primary data, whatever the potential benefits, and also lack the funds to pay specialist market research agencies to gather and manage such data for them. In these cases, organizations may be forced to rely on data that is less than ‘perfect’ but that can be accessed more cheaply, e.g., from secondary sources

* Les contraintes budgétaires

1. Le SEM peut devenir coûteux, et les efforts connexes de collecte et de traitement des données peuvent être très onéreux. Certaines agences de publicité sont payées en pourcentage des dépenses publicitaires, ce qui ne les incite pas à maintenir les coûts publicitaires à un bas niveau, par exemple, même si les conversions sont élevées, le coût par conversion doit être minimisé.

2. De nombreuses organisations peuvent ne pas disposer de l'expertise nécessaire pour mener des enquêtes approfondies afin de recueillir des données primaires, quels que soient les avantages potentiels, et ne disposent pas non plus des fonds nécessaires pour payer des agences d'études de marché spécialisées pour recueillir et gérer ces données à leur place. Dans ces cas, les organisations peuvent être obligées de s'appuyer sur des données qui ne sont pas "parfaites" mais qui sont accessibles à moindre coût, par exemple à partir de sources secondaires

* Time constraints

Organizations are often forced to balance the need to build up as detailed a picture as possible regarding customer needs etc., against the desire to make decisions as quickly as possible. Fast decision-making is often a competitive advantage over bigger players where decisions require multiple stakeholders and take time to get approved.

* Contraintes de temps

Les organisations sont souvent obligées de trouver un équilibre entre la nécessité de dresser un tableau aussi détaillé que possible des besoins des clients, etc. et le désir de prendre des décisions le plus rapidement possible. La rapidité de la prise de décision est souvent un avantage concurrentiel par rapport aux grands acteurs lorsque les décisions nécessitent de multiples parties prenantes et prennent du temps à être approuvées.

* Reliability of the data

The value of any research findings depends critically on the accuracy of the data collected. Data quality can be compromised via a number of potential routes, e.g., leading questions, unrepresentative samples, biased interviewers, etc. Efforts to ensure that data is accurate, samples are representative and interviewers are objective will all add to the costs of the research but such costs are necessary if poor decisions and expensive mistakes are to be avoided.

* Fiabilité des données

La valeur de tout résultat de recherche dépend essentiellement de l'exactitude des données recueillies. La qualité des données peut être compromise par un certain nombre de voies possibles, par exemple des questions suggestives, des échantillons non représentatifs, des enquêteurs partiaux, etc. Les efforts déployés pour garantir l'exactitude des données, la représentativité des échantillons et l'objectivité des enquêteurs augmenteront les coûts de la recherche, mais ces coûts sont nécessaires si l'on veut éviter de mauvaises décisions et des erreurs coûteuses.

* Legal & ethical constraints

The Data Protection Act (1998) is a good example of a law that has a number of implications for market researchers collecting and holding personal data. For instance, researchers must ensure that the data they obtain is kept secure, is only used for lawful purposes, and is only kept for as long as it is necessary. It must be made clear as to why data is being collected and the consent of participants must be obtained. In addition to this, there are a number of guidelines, laid down by such organizations as the Market Research Society, that, although not legally binding, encourage organizations to behave ethically when dealing with members of the public.

* Contraintes juridiques et éthiques

La loi sur la protection des données (1998) est un bon exemple de loi qui a un certain nombre d'implications pour les études de marché qui collectent et détiennent des données personnelles. Par exemple, les chercheurs doivent s'assurer que les données qu'ils obtiennent sont conservées en toute sécurité, qu'elles ne sont utilisées qu'à des fins légales et qu'elles ne sont conservées qu'aussi longtemps que nécessaire. Il doit être clairement indiqué pourquoi les données sont collectées et le consentement des participants doit être obtenu. En outre, il existe un certain nombre de lignes directrices, établies par des organisations telles que la Market Research Society, qui, bien qu'elles ne soient pas juridiquement contraignantes, encouragent les organisations à se comporter de manière éthique lorsqu'elles traitent avec des membres du public.

In this instance the 2 most critical constraints are:

* Time – the domains will be sold soon and
* Money - a small marketing budget for search engine marketing

Dans ce cas, les deux contraintes les plus critiques sont :

* Le temps - les domaines seront bientôt vendus et
* L'argent - un petit budget pour le marketing des moteurs de recherche

**(6)  What data do we have?**

The company has offered no data of their own e.g., if they had an adwords account already, that would be a great place to check the data for what keywords have been working. You could even find traffic and cost estimates for the two keywords, ‘analytics’ and ‘data science’ to help with the decision.

You do have access to Google Trends data, you could start by checking if the CEOs assertion is correct!

(6) Quelles sont les données dont nous disposons ?

La société n'a pas proposé de données propres, par exemple, si elle avait déjà un compte adwords, ce serait un bon endroit pour vérifier les données pour savoir quels mots clés ont fonctionné. Vous pourriez même trouver des estimations de trafic et de coûts pour les deux mots clés "analytics" et "data science" pour vous aider à prendre une décision.

Vous avez accès aux données de Google Trends, vous pourriez commencer par vérifier si l'affirmation du PDG est correcte !

**(7)  What data do we need to get?**

For this example, we will proceed with the free Google Trends data only however, in practice, we could consider other sources of data, proprietary sources, open data, survey data, or our own testing data.

(7) Quelles données devons-nous obtenir ?

Pour cet exemple, nous utiliserons uniquement les données gratuites de Google Trends. Toutefois, en pratique, nous pourrions envisager d'autres sources de données, des sources propriétaires, des données ouvertes, des données d'enquête ou nos propres données de test.

**Our analysis will proceed as follows:**

Notre analyse se déroulera comme suit :

**Step 1:**

**Using Google Trends data, start by checking if the CEO’s assertion is correct, see if ‘Analytics’ is the preferred search term.**

Étape 1 :

En utilisant les données de Google Trends, commencez par vérifier si l'affirmation du PDG est correcte, voyez si "Analytics" est le terme de recherche préféré.

**Step 2:**

**Using Google Trends data, we would check whether there are any other search terms relating to A and B that can help us answer the heart of the question, ‘which search term is better’.**

**We may define ‘better’ as, ‘which search term will speak directly to our target audience of machine learning enthusiasts and data scientists!'**

Étape 2 :

À l'aide des données de Google Trends, nous vérifions s'il existe d'autres termes de recherche relatifs à A et B qui peuvent nous aider à répondre au cœur de la question "quel est le meilleur terme de recherche".

Nous pouvons définir le terme "meilleur" comme "le terme de recherche qui s'adresse directement à notre public cible, les passionnés d'apprentissage machine et les scientifiques des données".

### CEO vs. CMO Case Study



Hello and welcome to Big Data University.

Today we're taking the course digital analytics and regression.

Aand this course is part of the data science for business learning path which is designed to give working professionals some practical examples about how they can use data science in the workplace.

Bonjour et bienvenue à la Big Data University.

Aujourd'hui, nous suivons le cours d'analyse et de régression numériques.

Et ce cours fait partie du parcours d'apprentissage de la science des données pour les entreprises, qui est conçu pour donner aux professionnels en activité quelques exemples pratiques sur la façon dont ils peuvent utiliser la science des données sur le lieu de travail.



Most companies all companies have data nowadays big medium and small.

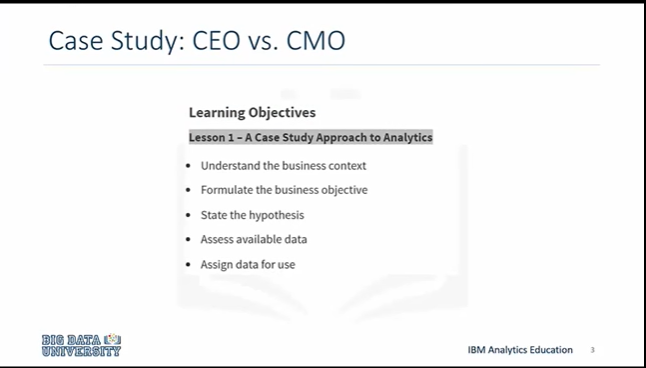
However we are in the process of unlocking the value from that data so we can really start to see results.

Let's get started.

La plupart des entreprises toutes les entreprises ont des données aujourd'hui grandes, moyennes et petites.

Cependant, nous sommes en train de libérer la valeur de ces données afin de pouvoir commencer à voir des résultats.

Commençons.

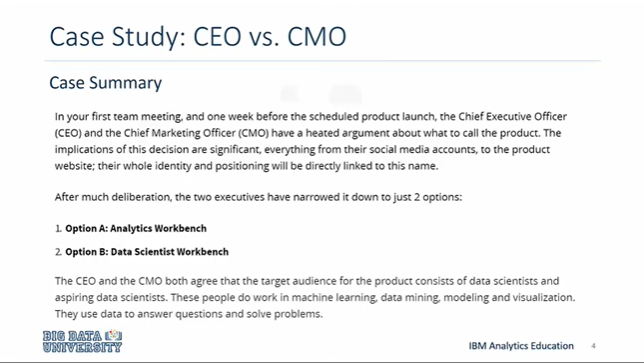


Today we're going to go through a case study in this case we'll be looking at the following learning objectives.

* Firstly we need to understand the business context.
* Secondly we'll formulate a business objective.
* Then we'll put down our hypothesis,
* we will assess the available data
* and then we'll assign data to use to answer the case.

Aujourd'hui, nous allons passer en revue une étude de cas. Dans ce cas, nous allons nous pencher sur les objectifs d'apprentissage suivants.

* Tout d'abord, nous devons comprendre le contexte de l'entreprise.
* Ensuite, nous formulerons un objectif commercial.
* Ensuite, nous formulerons notre hypothèse,
* nous évaluerons les données disponibles
* et ensuite nous assignerons les données à utiliser pour répondre à l'affaire.

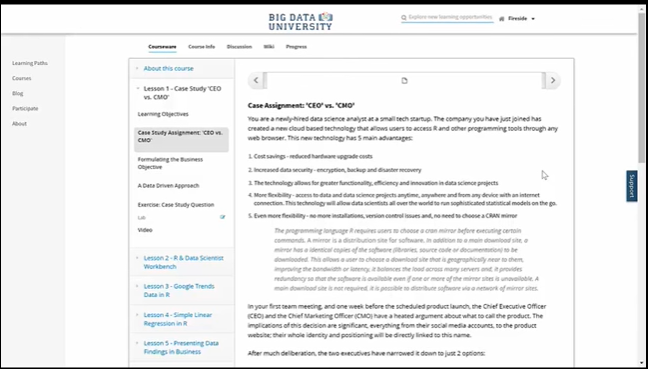


Let's look at the case study.

You have just joined a start-up that has created a fantastic product that allows data scientists all over the world to access data science tools and programming languages through their web browser.

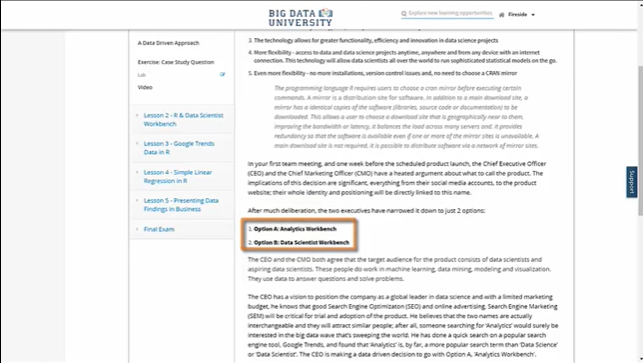
Examinons l'étude de cas.

Vous venez de rejoindre une start-up qui a créé un produit fantastique qui permet aux scientifiques du monde entier d'accéder à des outils et à des langages de programmation en sciences des données par le biais de leur navigateur web.



The big debate in the company at the moment is between the chief executive officer the CEO and the chief marketing officer the CMO.

Le grand débat dans l'entreprise en ce moment est entre le PDG le PDG et le directeur du marketing le CMO.

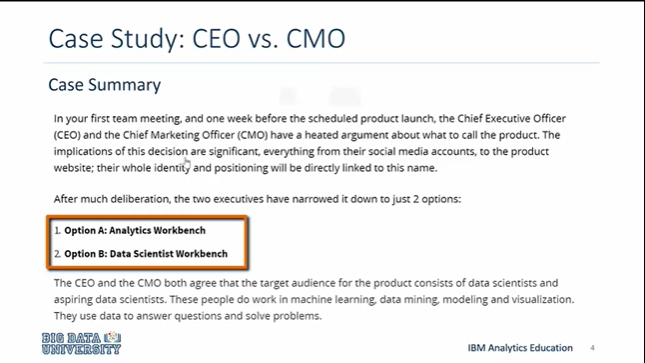


The discussion is as follows what to name this product :

* option a is analytics workbench,
* option B is data scientist workbench.

La discussion porte sur le nom à donner à ce produit :

* l'option a est l'Analytics Workbench,
* l'option B est l'atelier des scientifiques des données.



The CEO has already done some work, he has looked at Google Trends and he has found that analytics has by far many more searches than data scientists workbench he believes he's making a data-driven decision by choosing to call the product analytics workbench.

The CMO however has had some experience in search engine marketing and search engine optimization.

He has a gut feeling that analytics is too broad a search term and in the long run it's going to be very expensive for them when many people click on it who are not looking for data science related content.

He believes that data scientists workbench is a better name because it accurately targets the right group.

The right group consists of people looking for content relating to machine learning, data visualization, modeling and data mining.

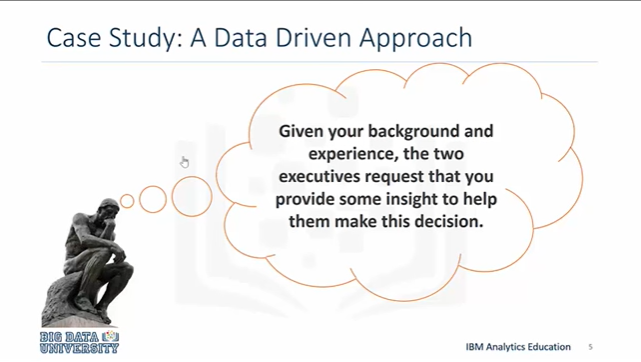
Le PDG a déjà fait quelques travaux, il a regardé Google Trends et il a trouvé que l'analyse a beaucoup plus de recherches que l'établi des scientifiques des données ; il pense qu'il prend une décision basée sur les données en choisissant d'appeler l'établi d'analyse des produits.

Le CMO a cependant une certaine expérience du marketing et de l'optimisation des moteurs de recherche.

Il a l'intuition que l'analyse est un terme de recherche trop large et qu'à long terme, elle va leur coûter très cher lorsque de nombreuses personnes cliqueront dessus sans chercher de contenu lié aux sciences des données.

Il pense que Data Scientific Workbench est un meilleur nom parce qu'il cible précisément le bon groupe.

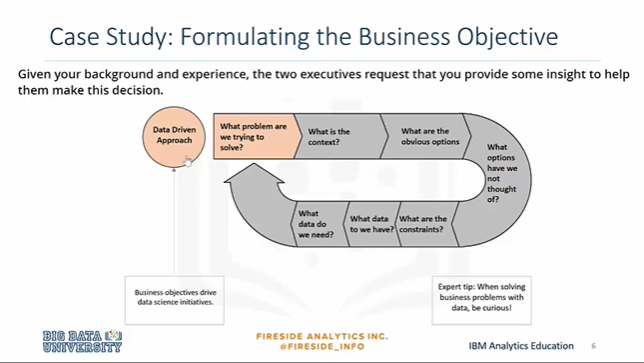
Le bon groupe est constitué de personnes qui recherchent des contenus relatifs à l'apprentissage machine, à la visualisation de données, à la modélisation et à l'exploration de données.



Given your background and experience the two executives request that you provide some insight to help

them make this decision.

Compte tenu de vos antécédents et de votre expérience, les deux dirigeants vous demandent de leur donner quelques indications pour les aider à prendre cette décision.



What does it mean to be data-driven ?

Fire site analytics provides a seven step process towards making a data-driven decision.

This process begins by understanding the business objective.

What problem are we trying to solve ?

What key decision needs to be made ?

Next we ask what's the context are they any noteworthy people dynamics ?

Because it's a start-up let's run with the assumption that this is a flat organization and because the two have asked you for your opinion you should actually give it to them.

Next we ask what are the obvious options ?

Well in this scenario we are given two options only a or B.

Next we ask what haven't we considered ?

Again we are not invited to explore different search terms that could make even better names that the two we have on the table.

The decision we must make is between one or the other.

What are the constraints ?

The usual time and money.

In this case study the domains will be sold soon and money a small marketing budget means that the organization must be very lean with their spending.

Next we ask what data do we have ?

Well the CEO has already given us a clue by telling us he's done an analysis in Google Trends.

The first thing we can do is verify whether his analysis is correct.

Next what data do we need to get ?

The CMO talks about how he would have liked to conduct surveys or a/b testing in order to gain more insights. However we're going to work with what we have Google Trends data.

From asking these simple questions we are now able to formulate a plan of how to approach the case problem ?

Or analysis will proceed as follows :

* step one we're going to start by checking if the CEOs assertion is correct. Is it correct to say that analytics has by far more searches than data scientists workbench ?
* step number two, which search term is better. Are there other keywords that we can use to test our keyword ?

Que signifie être axé sur les données ?

L'analyse des sites d'incendie fournit un processus en sept étapes pour prendre une décision fondée sur des données.

Ce processus commence par la compréhension de l'objectif commercial.

Quel problème essayons-nous de résoudre ?

Quelle décision clé doit être prise ?

Ensuite, nous nous demandons dans quel contexte se situe la dynamique des personnes ?

Comme il s'agit d'une start-up, partons du principe qu'il s'agit d'une organisation plate et que les deux parties vous ont demandé votre avis, vous devriez en fait le leur donner.

Ensuite, nous vous demandons quelles sont les options évidentes ?

Dans ce scénario, nous avons deux options, seulement a ou B.

Ensuite, nous vous demandons ce que nous n'avons pas envisagé.

Là encore, nous ne sommes pas invités à explorer les différents termes de recherche qui pourraient faire des noms encore meilleurs que ceux des deux que nous avons sur la table.

La décision que nous devons prendre est entre l'un ou l'autre.

Quelles sont les contraintes ?

Le temps et l'argent habituels.

Dans cette étude de cas, les domaines seront bientôt vendus et l'argent ; un petit budget de marketing signifie que l'organisation doit être très légère dans ses dépenses.

Ensuite, nous demandons quelles sont les données dont nous disposons.

Le PDG nous a déjà donné un indice en nous disant qu'il avait fait une analyse dans Google Trends.

La première chose que nous pouvons faire est de vérifier si son analyse est correcte.

Ensuite, quelles données devons-nous obtenir ?

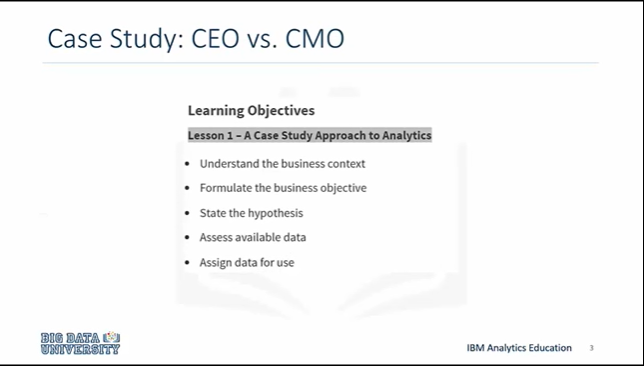
L'OCM nous dit qu'il aurait aimé mener des enquêtes ou des tests a/b afin d'obtenir plus d'informations. Cependant, nous allons travailler avec les données de Google Trends que nous avons.

En posant ces questions simples, nous sommes maintenant en mesure de formuler un plan d'approche du problème ?

Ou bien l'analyse se déroulera comme suit :

- à la première étape, nous allons commencer par vérifier si l'affirmation du PDG est correcte. Est-il exact de dire que l'analytique a beaucoup plus de recherches que l'atelier des scientifiques ?

- étape numéro deux, quel est le meilleur terme de recherche. Y a-t-il d'autres mots-clés que nous pouvons utiliser pour tester notre mot-clé ?



So we've understood the salient points in the case.

We have two options to choose from we've decided on a data set to use and we've decided on an approach.

We have in hypothesis something to prove or disprove.

Is analytics the more highly ranked search term ?

Secondly is analytics the better search term relative to the target audience we want ?

Read through the case material in detail and proceed to answer the assignment questions.

Nous avons donc compris les points saillants de l'affaire.

Nous avons le choix entre deux options : nous avons décidé d'un ensemble de données à utiliser et nous avons décidé d'une approche.

Nous avons dans l'hypothèse quelque chose à prouver ou à réfuter.

L'analyse est-elle le terme de recherche le mieux classé ?

Deuxièmement, l'analyse est-elle le meilleur terme de recherche par rapport au public cible que nous voulons ?

Lisez le dossier en détail et répondez aux questions de la mission.

### Module 2: Importing Google Trends data in R

**Learning Objectives**

**Lesson – Importing Google Trends with R**

* What is R?
* Loading Google Trends data into R
* Viewing the first rows of data

Leçon - Importer les tendances de Google avec R

* Qu'est-ce que la R ?
* Chargement des données de Google Trends dans R
* Visualisation des premières lignes de données

## IBM Developer Skills Network-Labs

### ****IBM Developer Skills Network - Labs****

Réseau de compétences des développeurs IBM – Laboratoires

IBM Developer Skills Network - Labs is a free cloud-based platform that allows users to access programming software through their browser.  This powerful technology means that you can access your work wherever you are, anytime, anywhere, and with any device; all you need is an internet connection and a data set. Data scientists can now access powerful analysis tools from home to office, coffee shop to internet café, from Argentina to Zimbabwe.

IBM Developer Skills Network - Labs est une plateforme gratuite basée sur le cloud qui permet aux utilisateurs d'accéder à des logiciels de programmation par le biais de leur navigateur. Cette puissante technologie signifie que vous pouvez accéder à votre travail où que vous soyez, à tout moment, n'importe où et avec n'importe quel appareil ; tout ce dont vous avez besoin est une connexion Internet et un ensemble de données. Les spécialistes des données peuvent désormais accéder à de puissants outils d'analyse de la maison au bureau, du café au café Internet, de l'Argentine au Zimbabwe.

The IBM Developer Skills Network - Labs enables data scientists to explore, prepare, refine, analyze, and visualize big-data interactively and collaboratively using a variety of open-source tools. The workbench is a web-based walk-up cloud environment with tools like OpenRefine, Jupyter Notebooks, Zeppelin Notebooks, R Studio IDE with easy access to Spark, Hadoop, dashDB, Cloudant, and a variety of other big data analytic engines, empowering data scientists to be more productive on their own terms.

Le réseau IBM Developer Skills Network - Labs permet aux spécialistes des données d'explorer, de préparer, d'affiner, d'analyser et de visualiser des données volumineuses de manière interactive et collaborative à l'aide de divers outils à source ouverte. Le workbench est un environnement en nuage basé sur le web avec des outils comme OpenRefine, Jupyter Notebooks, Zeppelin Notebooks, R Studio IDE avec un accès facile à Spark, Hadoop, dashDB, Cloudant, et une variété d'autres moteurs d'analyse de données importantes, permettant aux scientifiques d'être plus productifs selon leurs propres conditions.

In your preferred browser, go to [https://labs.cognitiveclass.ai/](http://www.datascientistworkbench.com/) and register. If you already have an account, log in.

Dans votre navigateur préféré, allez sur https://labs.cognitiveclass.ai/ et enregistrez-vous. Si vous avez déjà un compte, connectez-vous.

### Supported Browsers:

1. Google Chrome 38+
2. Safari 6+
3. Firefox 32+

Navigateurs supportés :

1. Google Chrome 38+

2. Safari 6+

3. Firefox 32+

### NB: Unsupported Browsers - Internet Explorer

* Once you are logged into DSW, you will have access to a number of data science tools from programming languages like R and Python to notebooks.
* R and Python are programming languages that use command-line instructions entered by a user to perform different functions.
* The R or Python instructions in a 'script' may be followed and executed on a user's local computer however, instructions in this course will direct users to IBM Developer Skills Network - Labs.

NB : Navigateurs non pris en charge - Internet Explorer

* Une fois que vous êtes connecté à DSW, vous aurez accès à un certain nombre d'outils de science des données, des langages de programmation comme R et Python aux ordinateurs portables.
* R et Python sont des langages de programmation qui utilisent des instructions en ligne de commande entrées par un utilisateur pour exécuter différentes fonctions.
* Les instructions R ou Python d'un "script" peuvent être suivies et exécutées sur l'ordinateur local de l'utilisateur. Toutefois, les instructions de ce cours orienteront les utilisateurs vers le réseau de compétences des développeurs IBM - Labs.

**Instructions:**

1. Register for IBM Developer Skills Network - Labs at [https://labs.cognitiveclass.ai/](http://www.datascientistworkbench.com/)
2. Log in to IBM Developer Skills Network - Labs
3. Take a [video tour](https://datascientistworkbench.com/" \l "video-tour" \t "_blank) of IBM Developer Skills Network - Labs
4. **NB: The initial set up of IBM Developer Skills Network - Labs may take up to 30 minutes**

Instructions :

1. Inscrivez-vous au réseau IBM Developer Skills Network - Labs à l'adresse https://labs.cognitiveclass.ai/

2. Connectez-vous au réseau IBM Developer Skills Network - Labs

3. Faites une visite vidéo du réseau de compétences des développeurs IBM - Labs

4. NB : La mise en place initiale du réseau IBM Developer Skills Network - Labs peut prendre jusqu'à 30 minutes

## What is R

### What is R?

Qu'est-ce que la R ?

R is a programming language that has grown in popularity in recent times due to its ease of use, statistical application, and its ability to handle big data sets, big data! We will be using R to analyze data through CC Labs.

R est un langage de programmation qui a gagné en popularité ces derniers temps en raison de sa facilité d'utilisation, de son application statistique et de sa capacité à traiter de gros ensembles de données, de grosses données ! Nous utiliserons R pour analyser les données par le biais de CC Labs.

R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the R Development Core Team, of which Chambers is a member. R is named partly after the first names of the first two R authors in about 1993. It is a programming language that is used for data manipulation, visualization, and statistics. R is a command-line programming language and is free and open source. We will be using R Studio, an integrated development environment (IDE), a user interface that makes it easier to manage the project we will be working on.

R a été créé par Ross Ihaka et Robert Gentleman à l'université d'Auckland, en Nouvelle-Zélande, et est actuellement développé par l'équipe centrale de développement de R, dont Chambers est membre. R est nommé en partie d'après les prénoms des deux premiers auteurs de R vers 1993. C'est un langage de programmation qui est utilisé pour la manipulation de données, la visualisation et les statistiques. R est un langage de programmation en ligne de commande et est gratuit et à code source ouvert. Nous utiliserons R Studio, un environnement de développement intégré (IDE), une interface utilisateur qui facilite la gestion du projet sur lequel nous allons travailler.

### R vs. Traditional Spreadsheet Software

R vs. logiciel de tableur traditionnel

Many users will be familiar with working in spreadsheet programs such as Microsoft Excel. The good news is that all the same functions (and more) are available in R. R gives users more flexibility in terms of how they can access, import, manipulate, query, and analyze data, especially big data.

De nombreux utilisateurs sont habitués à travailler dans des tableurs tels que Microsoft Excel. La bonne nouvelle est que toutes les mêmes fonctions (et plus) sont disponibles dans R. R offre aux utilisateurs une plus grande flexibilité en termes d'accès, d'importation, de manipulation, d'interrogation et d'analyse des données, en particulier des données volumineuses.

### Reasons why R is better than Excel for data analyses:

Les raisons pour lesquelles R est meilleur qu'Excel pour les analyses de données :

• More powerful data manipulation capabilities  
• Easier automation  
• Faster computation  
• It reads any type of data  
• Easier project organization  
• It supports larger data sets  
• Reproducibility (important for detecting errors)  
• Easier to find and fix errors  
• It's free  
• It's open-source  
• Advanced Statistics capabilities  
• State-of-the-art graphics  
• It runs on many platforms  
• Anyone can contribute packages to improve its functionality

* Des capacités de manipulation de données plus puissantes
* Une automatisation plus facile
* Calculs plus rapides
* Il lit tout type de données
* Une organisation plus facile des projets
* Il prend en charge des ensembles de données plus importants
* Reproductibilité (important pour la détection des erreurs)
* Plus facile à trouver et à corriger les erreurs
* C'est gratuit
* C'est un logiciel libre
* Capacités statistiques avancées
* Graphiques de pointe
* Il fonctionne sur de nombreuses plateformes
* Tout le monde peut contribuer à l'amélioration de ses fonctionnalités

**\*\*\*Expert tip \*\*\***

\*\*\*Conseil d'expert \*\*\*

Working in R is like driving a manual car. Manual drivers can drive automatic cars but it's not so easy the other way around. Many many people drive automatic cars and they're great for most driving needs however, if you would like to be a race car driver, it is best to know more about cars. Many analysts run models with the push of a button however in data science, it is extremely helpful to know what the data equivalent of a carburetor is.

Travailler dans la R, c'est comme conduire une voiture manuelle. Les conducteurs manuels peuvent conduire des voitures automatiques, mais ce n'est pas aussi facile dans l'autre sens. Beaucoup de gens conduisent des voitures automatiques et elles répondent à la plupart des besoins de conduite ; cependant, si vous souhaitez être pilote de course, il est préférable d'en savoir plus sur les voitures. Beaucoup d'analystes font tourner des modèles en appuyant sur un bouton, mais en science des données, il est extrêmement utile de connaître l'équivalent en données d'un carburateur.

### Using R in your browser via CC Labs

Utilisation de R dans votre navigateur via CC Labs

In this course, we will be using R in CC Labs. CC Labs is a cloud-based tool that allows users to access R (and other programming languages) through a web browser.

Dans ce cours, nous utiliserons R dans les CC Labs. CC Labs est un outil basé sur le cloud qui permet aux utilisateurs d'accéder à R (et à d'autres langages de programmation) par le biais d'un navigateur web.

## Loading Data into R in IBM Developer Skills Network - Labs

### Loading data into R

Let’s get started!

Commençons !

Log in to [IBM Developer Skills Network - Labs](https://labs.cognitiveclass.ai/)! Once you are logged in, look for ‘[RStudio IDE](https://labs.cognitiveclass.ai/tools/rstudio-ide/" \t "_blank)’ and click on it. you can also use the below **LTI** in order to start 'RStudio'. 'RStudioIDE' appears in two places on the home page, on the left under 'Build Analytics' and in a box on the right, also under ‘Build Analytics’.

Connectez-vous au réseau IBM Developer Skills Network - Labs ! Une fois que vous êtes connecté, cherchez "RStudio IDE" et cliquez dessus. Vous pouvez également utiliser le LTI ci-dessous pour démarrer "RStudio". RStudioIDE" apparaît à deux endroits sur la page d'accueil, à gauche sous "Build Analytics" et dans une boîte à droite, également sous "Build Analytics".

It may be helpful to split your screen so that you can see these instructions in your current IBM Developer Skills Network window and the IBM Developer Skills Network - Lab window at the same time. Split your screen by holding down the 'Window' button and the 'Left Arrow' simultaneously. Then click on the IBM Developer Skills Network - Lab window and hold down the 'Window' button and the 'Right Arrow'. Try it.

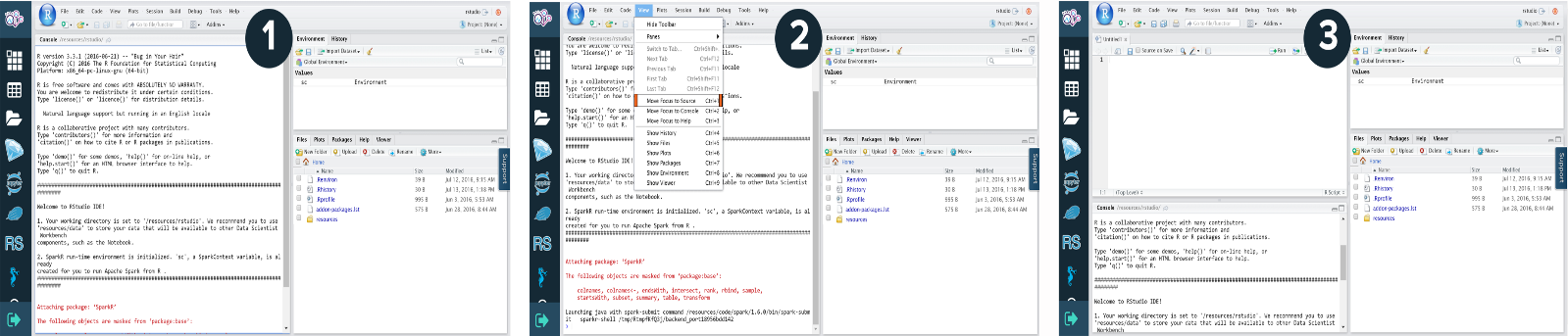
Il peut être utile de diviser votre écran pour que vous puissiez voir ces instructions dans votre fenêtre actuelle du IBM Developer Skills Network et dans la fenêtre du IBM Developer Skills Network - Lab en même temps. Divisez votre écran en maintenant simultanément le bouton "Fenêtre" et la "Flèche gauche" enfoncés. Cliquez ensuite sur la fenêtre du laboratoire du réseau de compétences des développeurs IBM et maintenez le bouton "Fenêtre" et la "Flèche droite" enfoncés. Essayez-le.

### Welcome to R Studio!

Bienvenue à R Studio !

Once you have clicked on '[RStudio IDE](https://labs.cognitiveclass.ai/tools/rstudio-ide/" \t "_blank)', you are now in a programming environment. When you open RStudio in IBM Developer Skills Network - Lab for the first time, the left half of the screen may be the 'Console' area. We will want to work in the RStudio 4 quadrant view. To retrieve the 'source' section of RStudio, use the shortcut keys Ctrl+1 or Command+1. You can achieve the same result by clicking on 'View' and 'Move Focus to Source'. We should now see the 4 quadrant view in RStudio. We can enter commands into the top left-hand quadrant and they will be executed in the bottom left-hand quadrant. Plots or charts will appear in your bottom right-hand quadrant.

Une fois que vous avez cliqué sur "RStudio IDE", vous êtes maintenant dans un environnement de programmation. Lorsque vous ouvrez RStudio dans IBM Developer Skills Network - Lab pour la première fois, la moitié gauche de l'écran peut être la zone "Console". Nous voudrons travailler dans la vue du quadrant 4 de RStudio. Pour récupérer la section "source" de RStudio, utilisez les touches de raccourci Ctrl+1 ou Command+1. Vous pouvez obtenir le même résultat en cliquant sur "View" et "Move Focus to Source". Nous devrions maintenant voir la vue en 4 quadrants dans RStudio. Nous pouvons entrer des commandes dans le quadrant supérieur gauche et elles seront exécutées dans le quadrant inférieur gauche. Des graphiques ou des diagrammes apparaîtront dans le quadrant inférieur droit.



Standard computer interface rules apply e.g., ‘File > Open’ opens files and you may use shortcuts. These are some popular RStudio shortcuts:

Ctrl + Enter: run the line of the script you are on

Ctrl + 1: Go to the source window, top left quadrant

Ctrl + 2: Go to the console window, bottom left quadrant

Ctrl + L: Clear the console, bottom left quadrant

Les règles d'interface informatique standard s'appliquent, par exemple, "Fichier > Ouvrir" ouvre les fichiers et vous pouvez utiliser des raccourcis. Il s'agit de certains raccourcis populaires de RStudio :

Ctrl + Entrée : lancez la ligne du script sur lequel vous vous trouvez

Ctrl + 1 : Aller à la fenêtre source, quadrant supérieur gauche

Ctrl + 2 : Aller à la fenêtre de la console, quadrant inférieur gauche

Ctrl + L : Effacer la console, quadrant inférieur gauche

To add data to your RStudio session, go to the ‘file’ icon in the top left of your screen. We will be using a script file with existing code and we would load it in the same way, go to the ‘file’ icon in the top left of your screen. The icon looks like an open file. You have the option to create a filing system to organize your uploads e.g., separate your data and your scripts. In the top right of that screen, you will see ‘Upload Data’.  Clicking on this will allow you to navigate to your local computer to upload the file you wish to work within R. In this instance, navigate to the ‘Regression File’ and click 'Upload'. Alternatively, find the file on your computer and simply drag and drop it into the ‘Upload Data’ box. Once you upload a file, it will be available for you to access in your cloud-based file directory. You need not 'set working directories' here, the files you drag and drop are in your R resources file.

Pour ajouter des données à votre session RStudio, allez sur l'icône "fichier" en haut à gauche de votre écran. Nous utiliserons un fichier script avec un code existant et nous le chargerons de la même manière, allez à l'icône "fichier" en haut à gauche de votre écran. L'icône ressemble à un fichier ouvert. Vous avez la possibilité de créer un système de classement pour organiser vos téléchargements, par exemple en séparant vos données et vos scripts. En haut à droite de cet écran, vous verrez "Télécharger des données". En cliquant dessus, vous pourrez vous rendre sur votre ordinateur local pour télécharger le fichier que vous souhaitez travailler dans R. Dans ce cas, allez dans le "Fichier de régression" et cliquez sur "Télécharger". Vous pouvez également trouver le fichier sur votre ordinateur et le glisser-déposer dans la case "Charger des données". Une fois que vous aurez téléchargé un fichier, vous pourrez y accéder dans votre répertoire de fichiers en mode "cloud". Vous n'avez pas besoin de "définir des répertoires de travail" ici, les fichiers que vous glissez et déposez sont dans votre fichier de ressources R.

After uploading a file, if we want to return to RStudio, we click on the RS icon on the right of the screen. The next time we click on, ‘File > Open’, we can open the file we just loaded and begin executing the code in it. If we have uploaded a data file like a CSV file, we can use the ‘Import Dataset’ option in the top right quadrant of R to bring that file into our RStudio session. A screen will pop up with a summary of what’s in the file and you can click ‘import’ to import that data into your RStudio session.

Après avoir téléchargé un fichier, si nous voulons retourner à RStudio, nous cliquons sur l'icône RS à droite de l'écran. La prochaine fois que nous cliquerons sur "Fichier > Ouvrir", nous pourrons ouvrir le fichier que nous venons de charger et commencer à exécuter le code qu'il contient. Si nous avons téléchargé un fichier de données comme un fichier CSV, nous pouvons utiliser l'option "Import Dataset" dans le quadrant supérieur droit de R pour amener ce fichier dans notre session RStudio. Un écran s'affichera avec un résumé du contenu du fichier et vous pourrez cliquer sur "importer" pour importer ces données dans votre session RStudio.

Instructions: Load this R script file [Loading\_Data\_into\_R\_Lesson\_2\_Script.R](https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-ML0103EN-CognitiveClass/labs/Digital_Analytics_Regression_Lessons.R) into IBM Developer Skills Network - Lab. Remember you can drag and drop this file from it's location on your computer to the 'upload' section in RStudio in IBM Developer Skills Network - Lab.

Instructions : Chargez ce fichier script R Chargement\_de\_données\_dans\_la\_leçon\_R\_2\_Script.R dans le IBM Developer Skills Network - Lab. N'oubliez pas que vous pouvez glisser et déposer ce fichier depuis son emplacement sur votre ordinateur vers la section "chargement" de RStudio dans IBM Developer Skills Network - Lab.

### Adding Data to RStudio in IBM Developer Skills Network - Lab

Ajout de données à RStudio dans le réseau IBM Developer Skills Network – Lab

We will be using a script file with existing code that you will work through to execute commands in RStudio. In order to access files in RStudio, we need to load them into IBM Developer Skills Network - Lab first.

Nous utiliserons un fichier script avec le code existant que vous utiliserez pour exécuter les commandes dans RStudio. Afin d'accéder aux fichiers dans RStudio, nous devons d'abord les charger dans IBM Developer Skills Network - Lab.

To add data to your IBM Developer Skills Network - Lab session, go to the ‘file’ icon in the top left of your screen. The icon looks like an open file. Here you have the option to create a filing system to organize your uploads e.g., separate your data and your scripts. In the top right of that screen you will see ‘Upload Data’.  Clicking on this will allow you to navigate to your local computer to upload the file you wish to work with in R. In this instance, navigate to the ‘Regression File’ and click 'Upload'. Alternatively, find the file in your computer and simply drag and drop it into the ‘Upload Data’ box. Once you upload a file, it will be available for you to access in your cloud based file directory. You will not need to 'set working directories' in IBM Developer Skills Network - Lab, your files are automatically stored in a resources in IBM Developer Skills Network - Lab. The files you drag and drop will be in your IBM Developer Skills Network - Lab resources folder.

Pour ajouter des données à votre session de laboratoire du réseau de compétences des développeurs IBM, cliquez sur l'icône "fichier" en haut à gauche de votre écran. L'icône ressemble à un fichier ouvert. Vous avez ici la possibilité de créer un système de classement pour organiser vos téléchargements, par exemple en séparant vos données et vos scripts. Dans le coin supérieur droit de cet écran, vous verrez "Télécharger des données". En cliquant dessus, vous pourrez vous rendre sur votre ordinateur local pour télécharger le fichier avec lequel vous souhaitez travailler dans R. Dans ce cas, allez au "Fichier de régression" et cliquez sur "Télécharger". Vous pouvez également trouver le fichier sur votre ordinateur et le glisser-déposer dans la case "Télécharger les données". Une fois que vous aurez téléchargé un fichier, vous pourrez y accéder dans votre répertoire de fichiers basé sur le cloud. Vous n'aurez pas besoin de "définir des répertoires de travail" dans IBM Developer Skills Network - Lab, vos fichiers sont automatiquement stockés dans une ressource dans IBM Developer Skills Network - Lab. Les fichiers que vous glisserez et déposerez se trouveront dans votre dossier de ressources d'IBM Developer Skills Network - Lab.

After uploading a file into IBM Developer Skills Network - Lab, we return to RStudio by clicking on the RS icon on the right of the screen. We are now back in the now familiar RStudio quadrant environment. The next time we click on, ‘File > Open’, we can open the file we just uploaded and begin executing the code in it.

Après avoir téléchargé un fichier dans IBM Developer Skills Network - Lab, nous retournons à RStudio en cliquant sur l'icône RS à droite de l'écran. Nous sommes maintenant de retour dans l'environnement désormais familier du quadrant RStudio. La prochaine fois que nous cliquerons sur "File > Open", nous pourrons ouvrir le fichier que nous venons de télécharger et commencer à exécuter le code qu'il contient.

If we have uploaded a data file like a CSV file, we can use the ‘Import Dataset’ option in the top right quadrant of R to bring that file into our RStudio session. A screen will pop up with a summary of what’s in the file and you can click ‘import’ to import that data into your RStudio session.

Si nous avons téléchargé un fichier de données comme un fichier CSV, nous pouvons utiliser l'option "Import Dataset" dans le quadrant supérieur droit de R pour amener ce fichier dans notre session RStudio. Un écran s'affichera avec un résumé du contenu du fichier et vous pourrez cliquer sur "importer" pour importer ces données dans votre session RStudio.

### Module 3: Plotting & Correlation

**Learning Objectives**

**Lesson 3 – Google Trends Data in R**

* Plotting Google Trends data in R
* Understanding the correlation between different terms searched on Google

Leçon 3 - Les données de Google Trends en R

* Tracer les données de Google Trends en R
* Comprendre la corrélation entre les différents termes recherchés sur Google

## Lab - Plotting Google Trends

### Accessing Google Trends in R

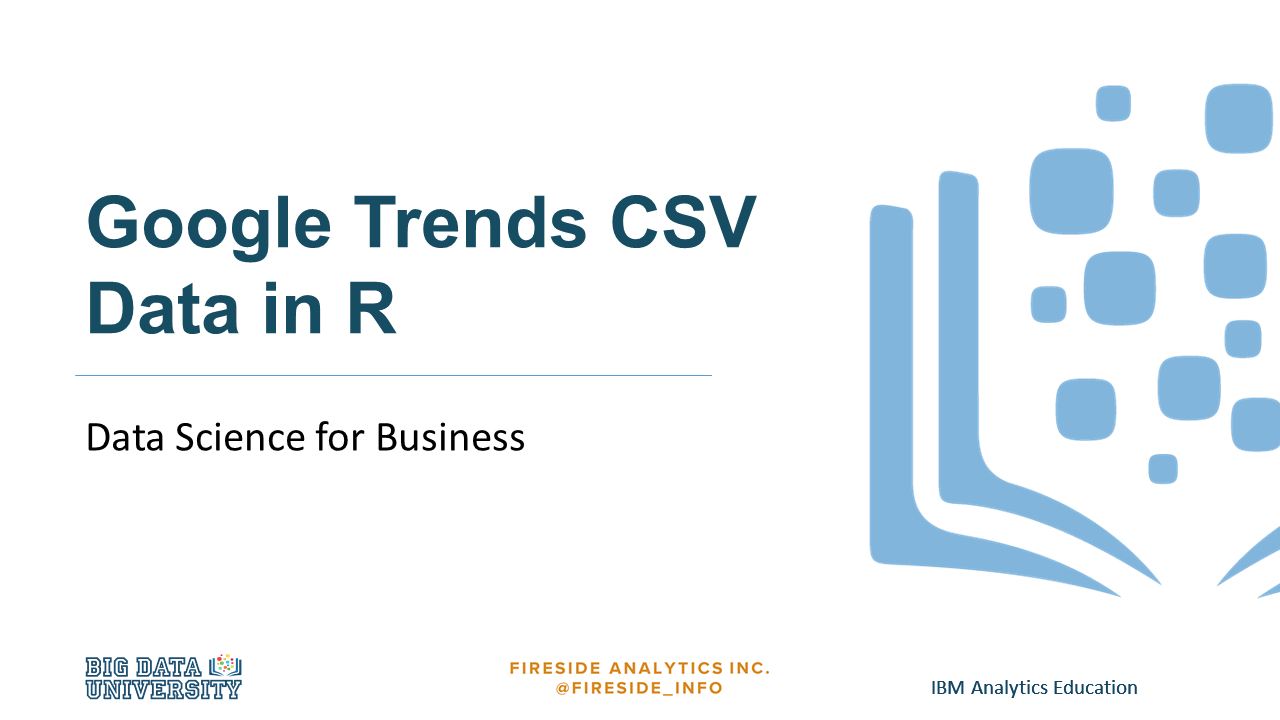
Accéder à Google Trends in R

Google Trends is a web tool that aggregates search terms and presents the search volume of different search terms relative to other search terms. If you are unfamiliar with it, you can see it at work here: <http://trends.google.com/> . Under 'Explore Topics', type in 'Analytics' and look at the result. In this lesson, we will access Google Trends data in RStudio and plot some search terms.

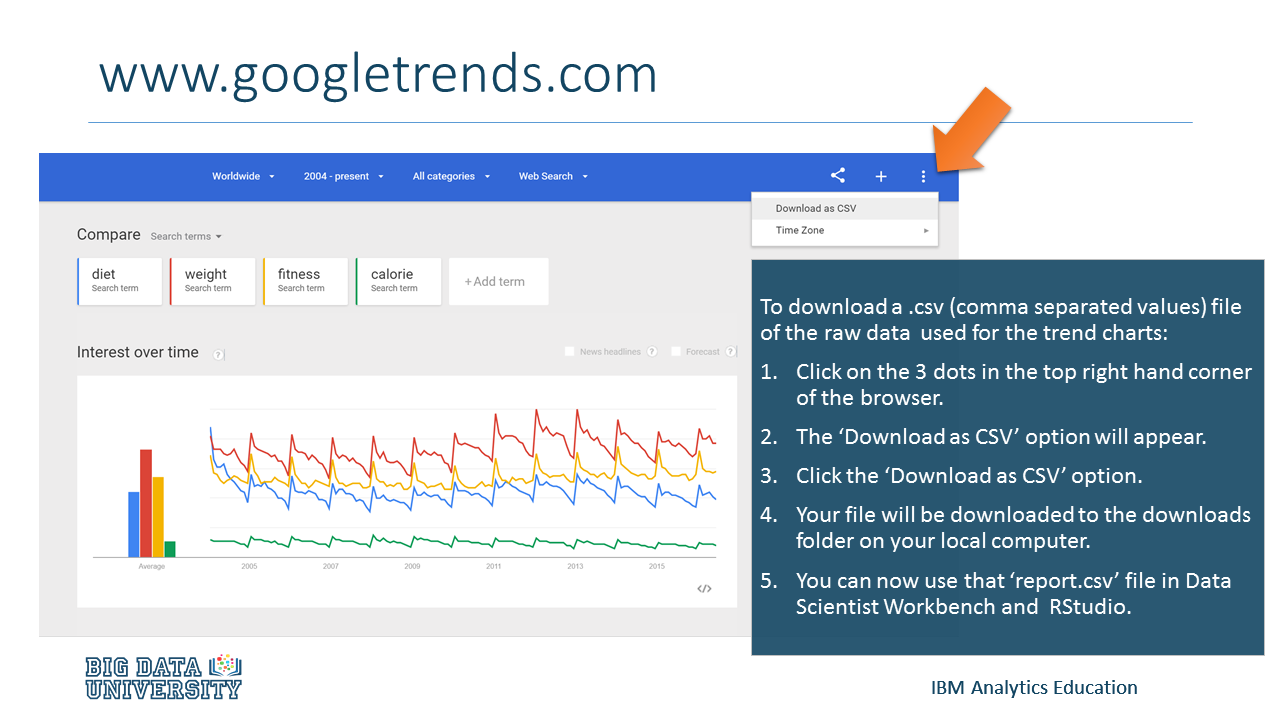
Google Trends est un outil web qui regroupe les termes de recherche et présente le volume de recherche des différents termes de recherche par rapport aux autres termes de recherche. Si vous ne le connaissez pas, vous pouvez le voir à l'œuvre ici : http://trends.google.com/ . Sous "Explore Topics", tapez "Analytics" et regardez le résultat. Dans cette leçon, nous allons accéder aux données de Google Trends dans RStudio et tracer quelques termes de recherche.

There is an R package called 'gtrendsR' that contains many lines of code or instructions to access Google Trends data within R. Using this package we would be able to pull Google Trends data directly into our RStudio session and plot trend charts. This package does however require users to input their Google login credentials in order to connect to Google Trends via RStudio. Warning: Your username and password for e.g., Gmail, would be used to access Google Trends data online and would be stored in your R history on Data Scientist Workbench. For this reason, we will rather download the Trends data '.csv file' first and upload that into Data Scientist Workbench for this lesson.

Il existe un package R appelé "gtrendsR" qui contient de nombreuses lignes de code ou d'instructions pour accéder aux données de Google Trends dans R. Grâce à ce package, nous pourrions tirer les données de Google Trends directement dans notre session RStudio et tracer des graphiques de tendances. Ce logiciel exige toutefois que les utilisateurs saisissent leurs identifiants de connexion Google afin de se connecter à Google Trends via RStudio. Attention : Votre nom d'utilisateur et votre mot de passe pour, par exemple, Gmail, seront utilisés pour accéder aux données Google Trends en ligne et seront stockés dans votre historique R sur Data Scientist Workbench. Pour cette raison, nous préférons télécharger d'abord les données de Tendances en fichier " .csv " et les charger dans Data Scientist Workbench pour cette leçon.









### Module 4: Simple Linear Regression in R

**Learning Objectives**

**Lesson 4 – Simple Linear Regression in R**

* Regression and Google Trends Data in R
* Box Plots and Histograms in R
* Scatterplots & Lines of best fit in R
* Simple Linear Regression in R

Leçon 4 - Régression linéaire simple en R

* Données sur la régression et les tendances de Google en R
* Box Plots et Histogrammes en R
* Diagrammes de dispersion et lignes de la meilleure concordance dans la R
* Régression linéaire simple en R

## Regression: Accessing Google Trends in R

### Exercise: Accessing Google Trends in RStudio

1. Download the R script file and the lesson .csv file below to your computer.
   * + [Lesson Exercise\_Google Trends Data (report.csv)](https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-ML0103EN-CognitiveClass/labs/report.csv)
     + [R script file: (Lesson\_Exercise\_2\_Digital\_Analytics\_Regression.R)](https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-ML0103EN-CognitiveClass/labs/Lesson_Exercise_2_Digital_Analytics_Regression.R)
2. Open the R script file in RStudio using the below **LTI** and follow the instructions.
3. To 'run' or execute the script, you may use the shortcut keys 'Ctrl / Command + Enter' or you may click on the 'run' button in the top right hand side of the top left-hand quadrant in RStudio.
4. Read the notes in the R script file. Notes are preceded by the '#' sign in a command line.
5. Follow the theory lessons in Big Data University while running the R script.

### Exercise: R Script

1. Run the following lines of script in RStudio.



## Video

hello and welcome to Big Data University today we're going to be covering simple linear regression in R and we'll be

doing a statistics review our learningobjectives our box plots and histograms and our scatter plots and lines of best

fit in R and lastly simple linear regression in R let's get started box plots sometimes referred to as box and whisker plots are a visual way of summarizing basic characteristics of a data set box plots show the highest and lowest values that are not outliers the middle value and the values at the first quarter and third quarter mark outliers are shown as dots after the whiskers this gives us a simple way to quickly understand the spread of data and it's a great way to quickly compare two data sets let's have a look at some box plots in our studio and data scientist workbench to begin we must first log in next we click on my data which takes us to the upload page we'll upload our two files next we'll click on our studio once in our studio we'll adjust the view to the four quadrants that we used to then we'll open the our script file that we've just loaded looking through this our script file it begins with some notes and we know that these are notes because they begin with the hash the file we are looking at consists of Google Trends data now we are ready to look at a boxplot these are the lines of code you will need to run a boxplot by hitting ctrl or command enter on your keyboard the box plot appears in the bottom right hand quadrant of our studio let's look at another box plot histograms are a statistical way of representing the frequencies of data values and intervals a histogram is a chart plot of a frequency table where the height of the bars tells us how many data points are in each interval let's look at the first histogram let's look at the histograms for some highly correlated data the histograms looked the same scatter plots scatter plots are charts that visualize the relationship between two variables every data point for one variable is plotted against the corresponding data point for another variable and the resultant pattern looks like scattered data points line of best fit a line of best fit is a line drawn through a scatterplot so that

each point on that line minimizes the total distance to any of the scattered data points this is traditionally called

a least squares line and in this example it follows the formula y equals MX plus C this is the line of code you will need to run a scatterplot this is the line of code you will need to add a line of best fit correlation correlation is a numerical way of interpreting the relationship between two variables a regression analysis uses the least squares method to fit a line through a scatterplot and it's measured by r-squared the correlation coefficient is the square root of R squared taking on the sign plus or minus of the slope of the data when R is high and positive we say that there is a positive

correlation for example when X goes up Y goes up so what is regression we know that a line of best fit is a line drawn

through a scatter plot so that each point on that line minimizes the total distance to any of the scatter plot points in a scatter plot with the line of best fit running through it we can assess how well the X variable explains the change in the Y variable using regression for example how is the variation in sales explained by the variation in price when we find a high r-squared in percentage terms the changes on Y are largely explained by the changes in X if the r-squared is 95% for example we say that 95% of the variation in sales is explained by the variation in price and the rest the 5%

that's due to error the error referred to here is the distances between the dots and the scatter plot and the line of best fit to fit an ordinary linear model in r we use the LM function for linear model we assign the results to an

object code regression this is the line of code you will need to run a simple linear regression model in R to see the full regression output we use the summary function the full regression output includes a lot of useful information for example the p-value the F statistic and the r-squared we just learned about revise these concepts further

after the lesson we have reviewed some key statistical concepts and we've seen how to compute those concepts in our

happy learning

## Box Plots & Histograms in R

### What is a histogram?

Qu'est-ce qu'un histogramme ?

Histograms are a statistical way of representing the frequencies of data values at particular intervals. The more traditional description is that a histogram is a chart plot of a frequency table where the height of the bars in it tells us how many data points are in each interval.

Les histogrammes sont un moyen statistique de représenter les fréquences des valeurs des données à des intervalles particuliers. La description la plus traditionnelle est qu'un histogramme est un graphique d'un tableau de fréquences où la hauteur des barres nous indique combien de points de données se trouvent dans chaque intervalle.

### What is a box plot?

Qu'est-ce qu'un box plot ?

Box plots sometimes referred to as 'Box and Whisker Plots', are a visual way of summarizing basic characteristics of a data set. Box Plots show the highest and lowest values (that are not outliers), the middle value, and the values at the 1st quarter and 3rd quarter mark. Outliers are shown as dots after the 'whiskers'. This gives us a simple way to quickly understand the spread of the data and is great for quickly comparing two data sets.

Les diagrammes en boîtes, parfois appelés "diagrammes en boîtes et en moustaches", sont un moyen visuel de résumer les caractéristiques de base d'un ensemble de données. Les Box Plots montrent les valeurs les plus élevées et les plus basses (qui ne sont pas des valeurs aberrantes), la valeur moyenne, et les valeurs au 1er et au 3ème trimestre. Les valeurs aberrantes sont représentées par des points après les "moustaches". Cela nous permet de comprendre rapidement la répartition des données et de comparer rapidement deux ensembles de données.

Think of the quartiles as, if your data set was ranked from its lowest to its highest value, Q1 would be the middle of the low values, (below the median) and Q3 is the middle of the high values (above the median). Box Plots break out a data set into 4 sets, before Q1, Q1 to the Median (Q2), the Median to Q3 and, after Q3. The interquartile range is defined as Q3 - Q1.

Imaginez les quartiles comme suit : si votre ensemble de données était classé de sa valeur la plus basse à sa valeur la plus haute, Q1 serait le milieu des valeurs basses (en dessous de la médiane) et Q3 est le milieu des valeurs hautes (au-dessus de la médiane). Les Box Plots répartissent un ensemble de données en 4 ensembles, avant Q1, de Q1 à la médiane (Q2), de la médiane à Q3 et, après Q3. L'intervalle interquartile est défini comme suit : Q3 - Q1.

The lowest part of the box in the Box Plot is Q1, there is a line inside the box, the median, and the end of the box, Q3.

La partie inférieure de la case dans le Box Plot est Q1, il y a une ligne à l'intérieur de la case, la médiane, et la fin de la case, Q3.

Outliers are typically defined as 1.5 \* (Q3 - Q1) and if a data point is (1.5\* the interquartile range) away from Q1 and Q3 (the edges of the box), it is considered an outlier. The whiskers of the box plot are lines from each end of the box out to the farthest data point that is not an outlier.

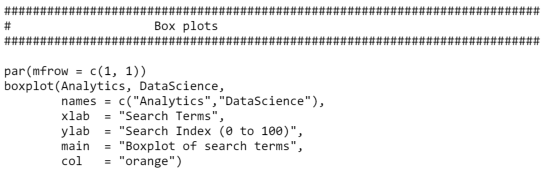
Les valeurs aberrantes sont généralement définies comme 1,5 \* (Q3 - Q1) et si un point de données est (1,5 \* l'intervalle interquartile) éloigné de Q1 et Q3 (les bords de la boîte), il est considéré comme une valeur aberrante. Les moustaches du box plot sont des lignes allant de chaque extrémité de la boîte jusqu'au point de données le plus éloigné qui n'est pas une valeur aberrante.

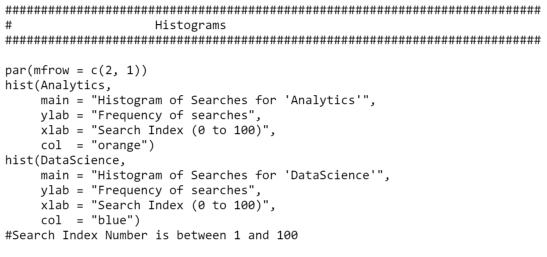
### Exercise: Accessing Google Trends in RStudio

1. Download the R script file and the lesson .csv file below to your computer.
   * + [Lesson Exercise\_Google Trends Data (report.csv)](https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-ML0103EN-CognitiveClass/labs/report.csv)
     + [R script file: (Lesson\_Exercise\_2\_Digital\_Analytics\_Regression.R)](https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-ML0103EN-CognitiveClass/labs/Lesson_Exercise_2_Digital_Analytics_Regression.R)
2. Open the R script file in RStudio using the below **LTI** and follow the instructions.
3. To 'run' or execute the script, you may use the shortcut keys 'Ctrl / Command + Enter' or you may click on the 'run' button in the top right hand side of the top left-hand quadrant in RStudio.
4. Read the notes in the R script file. Notes are preceded by the '#' sign in a command line.
5. Follow the theory lessons in Big Data University while running the R script.

### Exercise: R Script - Box Plots

1. Run the following lines of script in RStudio.





## Scatter Plots & Lines of Best Fit

### What is a Scatter Plot?

Qu'est-ce qu'un nuage de points ?

Scatter Plots are charts that visualize the relationship between two sets of data. Every data point for one variable is plotted against a corresponding data point for another variable and the resultant pattern looks like scattered data points.

Les diagrammes de dispersion sont des graphiques qui visualisent la relation entre deux ensembles de données. Chaque point de données d'une variable est tracé par rapport à un point de données correspondant d'une autre variable et le modèle qui en résulte ressemble à des points de données épars.

One of the most common relationships that can be shown in Scatter Plots is the relationship between a product's price and the number of units sold. Typically when the price goes down (a sale), more units are sold and when the price goes up, people buy less of a product. There are exceptions but that tends to be the case.

L'une des relations les plus courantes que l'on peut montrer dans les diagrammes de dispersion est la relation entre le prix d'un produit et le nombre d'unités vendues. Généralement, lorsque le prix baisse (une vente), plus d'unités sont vendues et lorsque le prix augmente, les gens achètent moins d'un produit. Il existe des exceptions, mais c'est généralement le cas.

The thing we want to reduce uncertainty about goes on the Y axis. We want to know what our sales will be if our price changes. The variable on the Y axis, sales, is the dependant variable. The variable on the X axis is the independent variable and it's the thing we can change, increase or decrease, to see what happens to Y.

La chose sur laquelle nous voulons réduire l'incertitude va sur l'axe Y. Nous voulons savoir quelles seront nos ventes si notre prix change. La variable sur l'axe Y, les ventes, est la variable dépendante. La variable sur l'axe X est la variable indépendante et c'est la chose que nous pouvons changer, augmenter ou diminuer, pour voir ce qui arrive à Y.

To plot a Scatter Plot, we take each value of Y and X for a particular observation e.g., a day, a week, a month and we mark them on a graph. Done.

Pour tracer un diagramme de dispersion, nous prenons chaque valeur de Y et X pour une observation particulière, par exemple un jour, une semaine, un mois, et nous les marquons sur un graphique. C'est fait.

A Scatter Plot showing many dots following the line between  11 and 5 on a clock is a negative linear relationship e.g., when price goes up, sales go down. A Scatter Plot running from 1 to 7 on a clock shows a positive relationship, when X goes up, so does Y.

Un diagramme de dispersion montrant de nombreux points suivant la ligne entre 11 et 5 sur une horloge est une relation linéaire négative, par exemple, lorsque le prix augmente, les ventes diminuent. Un nuage de points allant de 1 à 7 sur une horloge montre une relation positive, lorsque X augmente, Y augmente aussi.

### Line of Best Fit

Ligne de Best Fit

A line of best fit is a line drawn through a scatter plot so that each point on that line minimizes the total distance to any of the scatter data points. This is traditionally called a 'Least Squares Line' and it follows the formula, y = mX + C. Imagine a line running through a scatter plot. Each point on that line will have an X and a Y value. The least squares method says that we would take each line y value and subtract it from the scatter dots Y value. Our intention is to sum all these values however, some of the values in this subtraction will be negative as some scatter dots will be below the line so, we square each value before we sum all the values. The Line of Best Fit is one where that sum of the squares is the lowest number it can be i.e., all the dots are either on or very close to the line. When this happens, our line is a good way of predicting Y values, given values of X.

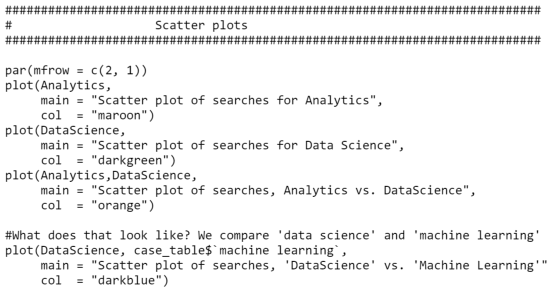
Une ligne de meilleur ajustement est une ligne tracée à travers un diagramme de dispersion de sorte que chaque point sur cette ligne minimise la distance totale à n'importe lequel des points de données de dispersion. Cette ligne est traditionnellement appelée "ligne des moindres carrés" et elle suit la formule y = mX + C. Imaginez une ligne passant par un nuage de points. Chaque point de cette ligne aura une valeur X et une valeur Y. Selon la méthode des moindres carrés, nous prenons la valeur y de chaque ligne et nous la soustrayons de la valeur Y des points de dispersion. Notre intention est d'additionner toutes ces valeurs, mais certaines des valeurs de cette soustraction seront négatives, car certains points de dispersion se trouveront en dessous de la ligne, de sorte que nous élèverons chaque valeur au carré avant d'additionner toutes les valeurs. La ligne de meilleur ajustement est celle où la somme des carrés est le plus petit nombre possible, c'est-à-dire que tous les points sont soit sur la ligne, soit très proches de celle-ci. Lorsque cela se produit, notre ligne est un bon moyen de prédire les valeurs Y, étant donné les valeurs X.

You are encouraged to revise this topic on your own. For our purposes, just note that the Best Fit Line can be used to predict values of Y, given values of X.

Vous êtes encouragé à réviser ce sujet par vous-même. Pour nos besoins, il suffit de noter que la ligne Best Fit peut être utilisée pour prédire les valeurs de Y, étant donné les valeurs de X.

### Exercise: Accessing Google Trends in RStudio

1. Download the R script file and the lesson .csv file below to your computer.
   * + [Lesson Exercise\_Google Trends Data (report.csv)](https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-ML0103EN-CognitiveClass/labs/report.csv)
     + [R script file: (Lesson\_Exercise\_2\_Digital\_Analytics\_Regression.R)](https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-ML0103EN-CognitiveClass/labs/Lesson_Exercise_2_Digital_Analytics_Regression.R)
2. Open the R script file in RStudio using the below **LTI** and follow the instructions.
3. To 'run' or execute the script, you may use the shortcut keys 'Ctrl / Command + Enter' or you may click on the 'run' button in the top right hand side of the top left-hand quadrant in RStudio.
4. Read the notes in the R script file. Notes are preceded by the '#' sign in a command line.
5. Follow the theory lessons in Big Data University while running the R script.



## Simple Linear Regression in R

### What is Regression?

Qu'est-ce que la régression ?

In a scatter plot with a line of best fit running through it, we can assess how well the X variable explains the changes in the Y variable using Regression e.g., What percentage of Sales is explained by changes in Price?

Dans un diagramme de dispersion parcouru par une ligne de meilleur ajustement, nous pouvons évaluer dans quelle mesure la variable X explique les changements de la variable Y en utilisant la régression, par exemple : Quel pourcentage des ventes est expliqué par les changements de prix ?

When we find a high 'R Squared', in percentage terms, the changes in Y are largely explained by the changes in X. If the R Squared is 95%, we say that, 95% of the variation in Sales is explained by the change in Price and the rest, the 5%, is due to error. The error referred to here is the distances between the dots in the scatter and the line of best fit.

Lorsque nous trouvons un "R au carré" élevé, en termes de pourcentage, les variations de Y sont largement expliquées par les variations de X. Si le R au carré est de 95%, nous disons que 95% de la variation des ventes est expliquée par la variation du prix et le reste, les 5%, est dû à une erreur. L'erreur dont il est question ici est la distance entre les points de la dispersion et la ligne du meilleur ajustement.

There are many other components to a Regression table output including, confidence intervals, P Values and t ratios. For a better understanding how to interpret Regression Models, you are strongly encouraged to revise this topic on your own.

La sortie d'un tableau de régression comporte de nombreux autres éléments, notamment des intervalles de confiance, des valeurs P et des rapports t. Pour mieux comprendre comment interpréter les modèles de régression, vous êtes fortement encouragé à réviser ce sujet par vous-même.

### What is Correlation?

Qu'est-ce que la corrélation ?

Correlation is a numerical way of interpreting the relationship between two variables. A Regression analysis uses the 'least squares method' to fit a line through a scatter plot and is measured by R Squared.

La corrélation est une façon numérique d'interpréter la relation entre deux variables. Une analyse de régression utilise la méthode des "moindres carrés" pour faire passer une ligne à travers un diagramme de dispersion et est mesurée par R au carré.

The Correlation coefficient is the square route of R Squared, taking on the sign (+ or -) of the slope of the data. When R is high and positive, we say that there is a positive correlation e.g., when X goes up, Y goes up.

Le coefficient de corrélation est le carré de R au carré, prenant le signe (+ ou -) de la pente des données. Lorsque R est élevé et positif, on dit qu'il y a une corrélation positive, par exemple, lorsque X augmente, Y augmente.

Correlation or r, measures the tightness of the scatter dots to the line of best fit and its sign tells us whether Y goes up with changes in X or Y goes down with changes in X. An r of 0 (zero) means that there is no relationship between X and Y and when r is 1, that means that there is a perfect relationship between X and Y where, when X goes up, Y goes up.

La corrélation ou r, mesure l'étroitesse des points de dispersion par rapport à la ligne du meilleur ajustement et son signe nous indique si Y augmente avec les changements dans X ou si Y diminue avec les changements dans X. Un r de 0 (zéro) signifie qu'il n'y a pas de relation entre X et Y et lorsque r est 1, cela signifie qu'il y a une relation parfaite entre X et Y où, lorsque X augmente, Y augmente.

**Recap:** A line of best fit is a line drawn through a scatter plot so that each point on that line minimizes the total distance to any of the scatter data points. This is traditionally called a 'Least Squares Line' and it follows the formula, y = mX + C. Imagine a line running through a scatter plot. Each point on that line will have an X and a Y value. The least squares method says that we would take each line y value and subtract it from the scatter dots Y value. Our intention is to sum all these values however, some of the values in this subtraction will be negative as some scatter dots will be below the line so, we square each value before we sum all the values. The Line of Best Fit is one where that sum of the squares is the lowest number it can be i.e., all the dots are either on or very close to the line. This summation is called 'Sum of Squared Errors'. When the SSE is at its lowest, our line is a good way of predicting Y values, given values of X.

Récapitulation : Une ligne de meilleur ajustement est une ligne tracée à travers un nuage de points de manière à ce que chaque point de cette ligne minimise la distance totale à l'un des points de données de diffusion. Cette ligne est traditionnellement appelée "ligne des moindres carrés" et elle suit la formule y = mX + C. Imaginez une ligne passant par un nuage de points. Chaque point de cette ligne aura une valeur X et une valeur Y. Selon la méthode des moindres carrés, nous prenons la valeur y de chaque ligne et nous la soustrayons de la valeur Y des points de dispersion. Notre intention est d'additionner toutes ces valeurs, mais certaines des valeurs de cette soustraction seront négatives, car certains points de dispersion se trouveront en dessous de la ligne, de sorte que nous élèverons chaque valeur au carré avant d'additionner toutes les valeurs. La ligne de meilleur ajustement est celle où la somme des carrés est le plus petit nombre possible, c'est-à-dire que tous les points sont soit sur la ligne, soit très proches de celle-ci. Cette addition est appelée "somme des erreurs au carré". Lorsque l'ESS est au plus bas, notre ligne est un bon moyen de prédire les valeurs de Y, étant donné les valeurs de X.

You are encouraged to revise this topic on your own.

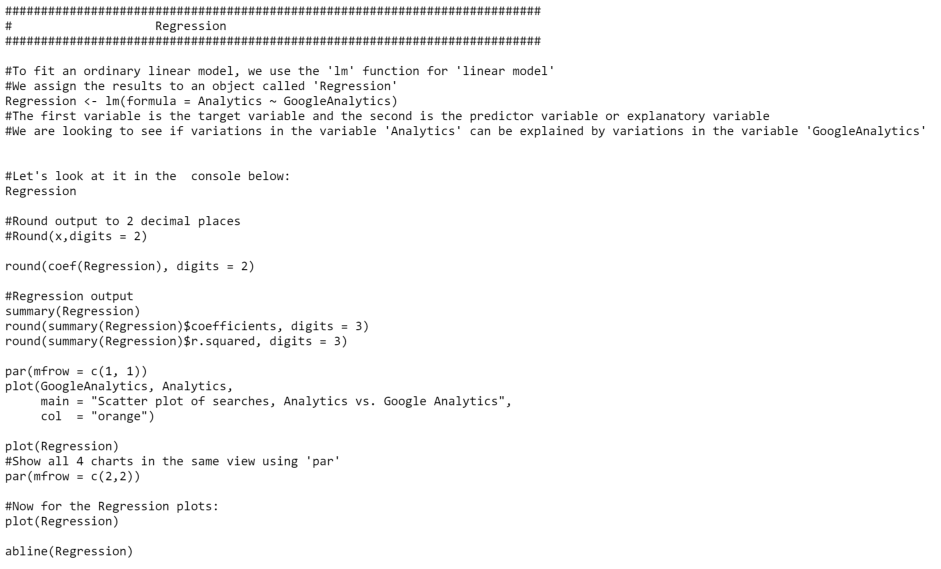
Vous êtes encouragé à réviser ce sujet de votre propre chef.

### Exercise: Accessing Google Trends in RStudio

1. Download the R script file and the lesson .csv file below to your computer.
   * + [Lesson Exercise\_Google Trends Data (report.csv)](https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-ML0103EN-CognitiveClass/labs/report.csv)
     + [R script file: (Lesson\_Exercise\_2\_Digital\_Analytics\_Regression.R)](https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-ML0103EN-CognitiveClass/labs/Lesson_Exercise_2_Digital_Analytics_Regression.R)
2. Open the R script file in RStudio using the below **LTI** and follow the instructions.
3. To 'run' or execute the script, you may use the shortcut keys 'Ctrl / Command + Enter' or you may click on the 'run' button in the top right hand side of the top left-hand quadrant in RStudio.
4. Read the notes in the R script file. Notes are preceded by the '#' sign in a command line.
5. Follow the theory lessons in Big Data University while running the R script.

### Exercise: R Script - Simple Linear Regression in R

1. Run the following lines of script in RStudio.



### Module 5: Presenting Data Findings in Business

### Learning Objectives

**Lesson 5 - Presenting Data Analytics in Business**

* Using data to answer a business question
* Summarizing the data analytics process
* Presenting data insights

Leçon 5 - Présentation de l'analyse des données dans les entreprises

* Utilisation de données pour répondre à une question commerciale
* Résumé du processus d'analyse des données
* Présentation des données

## Case Study Recap - CEO vs CMO

**Summary Case Study - CEO vs. CMO**

Résumé de l'étude de cas - CEO vs. CMO

You are a newly-hired data science analyst at a small tech startup. The company you have just joined has created a new cloud based technology that allows users to access R and other programming tools through any web browser. In your first team meeting, and one week before the scheduled product launch, the Chief Executive Officer (CEO) and the Chief Marketing Officer (CMO) have a heated argument about what to call the product.

Vous venez d'être embauché comme analyste en sciences des données dans une petite start-up technologique. La société que vous venez de rejoindre a créé une nouvelle technologie basée sur le cloud qui permet aux utilisateurs d'accéder à R et à d'autres outils de programmation par le biais de n'importe quel navigateur web. Lors de votre première réunion d'équipe, et une semaine avant le lancement prévu du produit, le directeur général (PDG) et le directeur du marketing (CMO) ont une discussion animée sur le nom à donner au produit.

Given your background and experience, the two executives request that you provide some insight to help them make this decision. They want to know which words to use in their product: "Analytics" or "Data Science" or "Data Scientist"

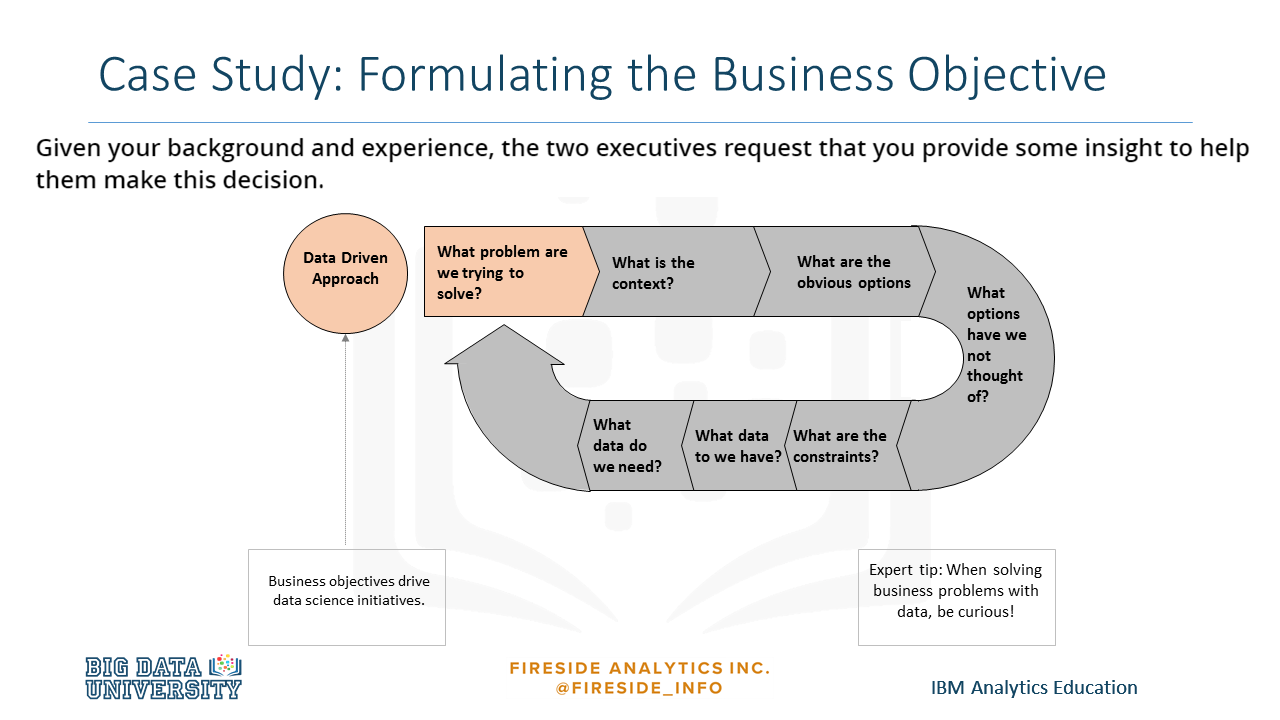
Compte tenu de vos antécédents et de votre expérience, les deux dirigeants vous demandent de leur fournir des informations pour les aider à prendre cette décision. Ils veulent savoir quels mots utiliser dans leur produit : "Analytique" ou "Science des données" ou "Data Scientist"

**Summary: Formulating the Business Objective**

Résumé : Formulation de l'objectif de l'entreprise

Developing the business objective for an analysis is a process of asking and answering a series of questions. Some of the questions we’ll need to ask ourselves in determining the business objectives are :

L'élaboration de l'objectif commercial d'une analyse est un processus consistant à poser et à répondre à une série de questions. Voici quelques-unes des questions que nous devrons nous poser pour déterminer les objectifs commerciaux :



**Summary: Analysis Approach**

Résumé : Approche analytique

Step 1:

Using Google Trends data, start by checking if the CEO’s assertion is correct, that ‘Analytics’ is the preferred search term.

Étape 1 :

En utilisant les données de Google Trends, commencez par vérifier si l'affirmation du PDG est correcte, à savoir que "Analytics" est le terme de recherche préféré.

Step 2:

Using Google Trends data, we would check whether there are any other search terms relating to "Analytics" that can help us answer the heart of the question, ‘which search term is better’.

We may define ‘better’ as, ‘which search term’ will speak directly to our target audience of machine learning enthusiasts and data scientists!

Étape 2 :

À l'aide des données de Google Trends, nous vérifions s'il existe d'autres termes de recherche relatifs à "Analytics" qui peuvent nous aider à répondre au cœur de la question, à savoir "quel terme de recherche est le meilleur".

Nous pouvons définir "meilleur" comme "quel terme de recherche" s'adressera directement à notre public cible, les passionnés d'apprentissage machine et les scientifiques des données !

## Presenting Data Insights

### Presenting Data Insights

Présentation des données

1. Through the steps in the data analytics process, you confirmed that the CEO’s assertion was correct! It is true that ‘Analytics’ had far more search terms than ‘Data Scientist’ however, you dug deeper into why searches for 'Analytics' were so high.
2. The Google Trends search index data showed that people searching for ‘Analytics’ were highly correlated with people searching for ‘Google Analytics’.  You inferred from this finding that if people searching for ‘Analytics’ were likely searching for something quite specific relating to Google advertising products and not general data science-related content.
3. You checked the correlation between searches for ‘Data Scientist’ and ‘Machine Learning’ and these two terms are highly correlated, whereas, the searches for ‘Data Scientist’ and ‘Analytics’ were not. This analysis led you to the conclusion that while searches for ‘Analytics’ were higher, based on the company’s objectives and desired target audience, the most appropriate product name is: ‘Data Scientist \_\_\_\_\_\_\_\_’. Given a consensus that it must contain the word "Workbench", the final name recommendation would be "Data Scientist Workbench".

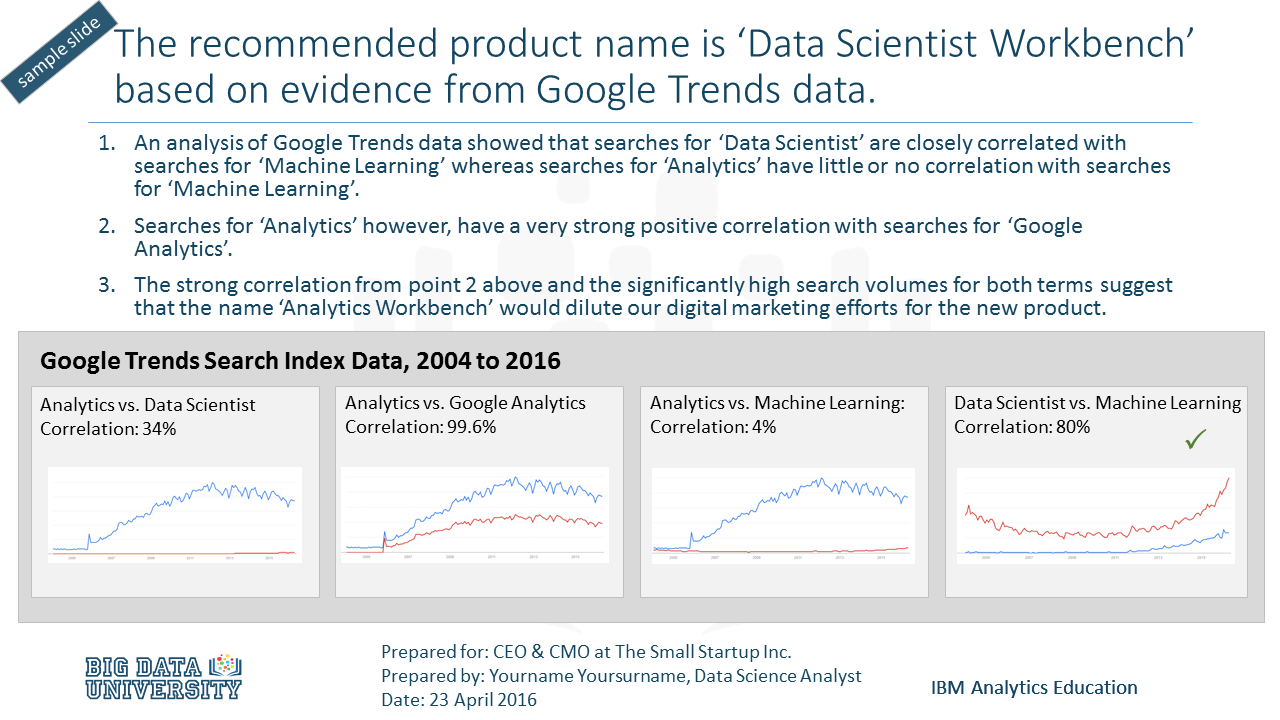
**1. Au fil des étapes du processus d'analyse des données, vous avez confirmé que l'affirmation du PDG était correcte ! Il est vrai que "Analytics" comportait beaucoup plus de termes de recherche que "Data Scientist", mais vous avez cherché à savoir pourquoi les recherches pour "Analytics" étaient si élevées.**

**2. Les données de l'index de recherche de Google Trends ont montré que les personnes recherchant "Analytics" étaient fortement corrélées avec les personnes recherchant "Google Analytics". Vous en avez déduit que si les personnes recherchant "Analytics" recherchaient probablement quelque chose de très spécifique en rapport avec les produits publicitaires de Google et non des données générales sur le contenu scientifique.**

**3. Vous avez vérifié la corrélation entre les recherches de "Data Scientist" et "Machine Learning" et ces deux termes sont fortement corrélés, alors que les recherches de "Data Scientist" et "Analytics" ne l'étaient pas. Cette analyse vous a amené à la conclusion que si les recherches pour "Analytics" étaient plus élevées, en fonction des objectifs de l'entreprise et du public cible souhaité, le nom de produit le plus approprié est : "Data Scientist \_\_\_\_\_\_\_\_". Compte tenu du consensus sur le fait qu'il doit contenir le mot "Workbench", la recommandation finale pour le nom serait "Data Scientist Workbench".**

**Your findings have been summarized in the sample slide below:**

Vos conclusions sont résumées dans l'exemple de diapositive ci-dessous :



## Presenting Data Insights

hello and welcome to this lesson in presenting data findings in business by fireside analytics for Big Data

University let's get started our learning objectives today are using data to answer a business question

summarizing the data analytics process and presenting data insights summarizing the data analytics process we begin by understanding the business objective business objectives drive data science initiatives and that is the data science approach we summarize the data analytics process by asking a number of questions in answering each of these questions we arrive at a data analytics approach to an analysis we start by asking what problem are we trying to solve then we ask what is the context what are the obvious options and what options have we not considered what are the constraints what data do we have and what data do we need telling a story with data is a three-step process we begin with the situation then we look at the complication and then we look at the resolution with data you can also think of this as the situation is the beginning the complication is the middle and the resolution is the end so what is the situation what is the beginning of the story well we describe the situation we analyze the current state what is happening what has happened we provide a context for the analysis when we look at the complication which is the middle of the story we consider what is the reason for the analysis what questions are we trying to answer what problem are we trying to solve lastly we look at the resolution and the end of the story how did we solve the problem with data what was the outcome what new situation are we in and what steps should we take next and we can tell the story with the data visualization a good data analytics presentation will follow this general flow however remember the CEO rule the CEO rule is as follows imagine that the CEO of a company you want to work with is attending your presentation she is a very busy and important person imagine that she suddenly gets up to take a phone call 15 minutes into your presentation make sure that the first 15 minutes of your presentation let her know that she must come back for the rest of it a good introduction or agenda slide in your presentation will help with this what does a good slide look like well it begins with the title and the title tells us why are we looking at his slide some good data visualization tips are charts must include an informative heading charts must have clearly labeled axes and data labels check the spelling and grammar in your charts use compelling colors that are easy to see use annotations to highlight theinteresting parts of your chart visualizations and data science presentations must pass the sowhat test that chart looks great but so what every chart and slide must add value to the audience otherwise it must be excluded from the presentation test whether or not to include a visualization by asking so what if the visualization does not meet at least one of the following five conditions excluded from the presentation one it shows new information - it is a demonstration of something three it is for emphasis 4 it is for re-emphasis and lastly it is a call to action if there is still uncertainty after the sowhat test remember be brief get to the point get to the most important points what do we mean by new information a new point something you wish to speak about in the presentation demonstration an existing point you've already spoken about but would like to show emphasis an existing point something you've already shown but would like to emphasize a great example of this is a chart and then a second chart with the same diagram where you zoom into a particular area re-emphasizing an existing point shown differently for example showing a table next to a chart with the same information different members in an audience understand information differently some people enjoy looking at the hard numbers others prefer charts

lastly a call to action what do you want the audience to do after seeing this visualization for example to realize something or to understand something here is a sample slide in this sample slide we start with a heading that tells us the so what of the slide this heading tells us why we are looking at the slide the bullet points provide logical evidence for the heading and the actual graphics the visualizations clearly and quickly tell the reader why the conclusion is the correct one this visualization groups four different smaller charts to make one graphic the overall chart is simple straightforward and to the point the chart also has some useful information at the bottom who it was prepared for who prepared it and the date you have now learned some tools to assist you in telling a compelling story

with data and you have learned some ways that you can use data visualization to make a great presentation happy learning