BIG DATA ANALYTICS MADE EASY PDF, EPUB, EBOOK

Y Lakshmi Prasad | 194 pages | 03 Dec 2016 | Notion Press, Inc. | 9781946390714 | English | none

(PDF) Download Big Data Analytics Made Easy - Y. Lakshmi Prasad - 1st Edition

Data Analytics Made Easy is an accessible guide to help you start analyzing data and quickly apply these skills to your work. The book introduces the concepts of data analytics and shows you how to get your data ready and apply ML algorithms. Create impressive visualizations with Microsoft Power BI and learn the greatest secret in successful analytics — how to tell a story with your data. By the end of this book, you will have learned how to implement machine learning algorithms and sell the results to your customers without writing a line of code. This creates friction when you have to switch between your desktop and the cloud just to reformat your data. But even if your data is clean, you still have to go through a lengthy publishing process to actually import your cloud-based data into Tableau Online.

For example, if you wanted to import your Google Analytics data, you would need to be familiar with SQL queries just to access it. Our brain has a natural bias for patterns —but we still need help connecting the dots with data. You just care about getting a quick chart. In this example for organic searches by date, we dragged over keywords to analyze traffic breakdown by keywords and fiscal quarter. If you want to do further analysis, you can drag and drop more data fields to the chart. If you want to analyze keyword breakdown by social network, for example, you can just drag over the Social Network field. You have to create a new widget and input several variables just to create a bar chart. You need an easy way to get this data to the rest of your team, instead of it staying siloed away. When your data changes, you need everyone on your team to be updated with the latest results so they can act on it.

We designed DataHero to end this problem of siloed data. It draws data from the applications your business already uses and brings it together into an all-in-one dashboard that you can customize and share. Your team can collaborate, your clients can access your dashboard, and everyone will see the latest results of your analysis. Your data is constantly changing, especially if you do several iterations of your keyword campaign, for example. Classification of analytics on the basis of business function and impact goes as follows: Marketing Analytics Sales and HR analytics Supply chain analytics and so on This can be a equitably long list as analytics has the prospective to impact virtually any business activity within a large organization.

But the most popular way of classifying analytics is o n the basis o f what it allo ws us to do. All the info r matio n is co llected differ ent industries and different departments. All we need to do is slicing and dicing the data in diverse ways, maybe looking at it from different angles or along different dimensions etc. As you can see descriptive analysis is possibly the simplest type of analytics to perform simply because it uses existing information from the past, to understand decisions in the present and hopefully helps decide an effective source of action in the future.

However, because of its relative ease of understanding and application descriptive analytics has been often considered the subdued twin of analytics. But it is also extremely powerful in its potential and in most business situations, Descriptive analytics can help address most problems. Retailers are very interested in understanding the relationship between products. They want to know if the person buys a product A, is he also likely buying product B or product C. This is called product affinity analysis or association analysis and it is commonly used in the retail industry. It is also called market basket analysis and is used to refer a set of techniques that can be applied to analyze the shopping basket or a transaction. Have you ever wondered why milk is placed right at the back of the store while magazines and chewing gum are right by the check-out? That is because through analytics retailers realize that while traveling all the way to the back of the store to pick up your essentials, you just may be tempted to pick up something else and also because magazines and chewing gum are cheap impulse buys.

You decide to throw them in your cart since they are not too expensive and you have probably been eying them as you waited in line at the counter. Predictive Analytics works by identifying patterns and historical data and then using statistics to make infer ences abo ut the futur e. At a ver y simplistic level, we tr y to fit the data into a certain pattern and if we believe the data is following a certain pattern then we can predict what will happen in the future. Do they prefer SMS or call numbers outside their city? This is information one can obtain purely by observation or descriptive analytics. But such companies would, more importantly, like to know which is the customers plan to leave and take a new connection with their competitors. This will use historical information but rely on predictive modeling and analysis to obtain results. This is predictive analysis. While descriptive analytics is a very powerful tool. A hotel owner wo uld want to pr edict ho w many o f his r o o ms will be o ccupied next week.

The CEO o f the Pharma Company will want to know which of his under test drugs is most likely to succeed. This is where predictive analytics is a lot more useful. In addition to these tools, there is a third type of analytics, which came into existence very recently, maybe just a decade old. This is called prescriptive analytics. Prescriptive analytics goes beyond predictive analytics by not only telling you what is going on but also what might happen and most importantly what to do about it.

It could also inform you about the impact of these decisions, which is what makes prescriptive analytics so cutting edge. Business domains that are great examples were prescriptive analytics can be used are the aviation industry or nationwide road networks. Prescriptive analytics can predict an effectively correct road bottlenecks, or identify roads where tolls can be implemented to streamline traffic. Airlines are always looking for ways to optimize their routes for maximum efficiency. This can be billions of dollars in savings but this is not that easy to do. So the aviatio n industr y o ften r elies o n pr escr iptive analytics to decide what, which and how they should fly their airplanes to keep cost down and profits up. So, we have taken a fairly in-depth look at descriptive, predictive and prescriptive analytics. The focus of this course is going to be descriptive analytics. Towards the end, we will also spend some time on understanding some of the more popular predictive modeling techniques.

Problem Identification: A problem is a situation that is judged as something that needs to be corrected. It is our job to make sure we are solving the right problem, it may not be the one presented to us by the client. What do we really need to solve? Sometimes the problem statements that we get from the business are very straight forward. For example: How do I identify the most valuable customers? How do I ensure that I minimize losses from the product not being available on the shelf? How do I optimize my inventory? How do I detect customers that are likely to default on a bill payment? These are straight forward problem statements and there is really no confusion around what is it that we are trying to achieve with an analytical project.

However, every single time our business statement may not lead to clear problem identification. Sometimes, the business statements are very high level and therefore you will need to spend time with the business to understand the needs and obtain the context. You may need to break down that issue into sub-issues to identify critical requirements. You may need to think about the constraints that need to be included in the solution. Let us take an example for this. Certainly, at a very high level, this is a valid business requirement. However, for your purpose which is to build a solution to address this question, is this a very valid statement or is it a sufficient starting point for the data analysis? Because, there are multiple problems with a business statement like this, which is, we want to receive credit care applications only from good customers. Let us look at the pr o blem with that pr o blem statement.

I want to r eceive cr edit car e applicatio ns o nly from good customers. One of the most obvious problem with that statement is who are good customers? That is, yo u spend o n the cr edit car d and yo u pay the cr edit car d co mpany back o n time. Why is that? These kinds of customers

are called revolvers. Who really is the good customer for a credit card company? Are these customers who pay on time? An answer could be both are good customers. How is that possible? It really depends on your perspective. They have a high revolving balance. Now, as an analyst, who decides who good customers are? When the credit card company gives you a business statement that says we want to accept credit card application from only good customers.

Do you know that they are looking at risk or revenue? It really depends on the business interest; it depends on the business goals for that year. In fact, a good customer this year may be a bad customer next year. This is why it is important to obtain the context or the problem statement before starting on an analysis. But this is not the only problem with this problem statement. Another problem is thinking about the decision which is, can you really insist on receiving g o o d applicatio ns o r can yo u insist o n appr o ving g o o d applicatio ns. Is the decisio n at the application stage or the approval stage? Can you really control applications to be good or can you control the decisions to enable only good customers to come on to you?. Another problem with this problem statement is that we only want to receive credit card applications from good customers. Is it realistic for you to assume that you will have a solution that will never accept a bad customer?

Again, not a realistic outcome. Coming back to our problem definition state which is, given a business problem, I want to get good customers as a credit card company. How do you frame that problem into something that analytical approach can tackle? One way is to add specifics to the problem statement. So, think about specific, measurable, attainable, realistic, and timely outcomes that you can attach to that problem statement.

That is, why we emphasize that you need to understand the business context thoroughly and talk to the business that you are tackling the right problem. How would I be able to add specifics to this pr o blem statement? Let us assume that I am lo o king at it fr o m the r isk per spective, because in this year my credit card companies focused on reducing the portfolio risk. So, I could have a variety of business problem statements. For example, reduce losses from cr edit car d default by at least 30 per cent in the fir st 12 mo nths po st implementatio n o f the new strategy.

Develop an algorithm to screen applications that do not meet good customer defined criteria that will reduce defaults by 20 percent in the next 3 months. Identify strategies to reduce defaults by 20 percent in the next three months by allowing at-risk customers additional payment options. We have decided that the good problem definition is something that we ar e tackling fr o m a r isk per spective. But, fo r the same business statement, we no w have three different problem statements that are tackling three different things. Again, which of these should I choose as a starting point for my analysis? Should I identify strategies for my existing customers or should I look at identifying potential new customers? Again, this is something that may be driven by business needs. So, it is important to constantly talk to the business to make sure that when you are starting an analytics project you are tackling the right problem statement. Getting to a clearly defined problem is often discovery driven — Start with a conceptual definition and through analysis root cause, impact analysis, etc.

A problem becomes known when a person observes a discrepancy between the way things are and the way things ought to be. Root Cause Analysis is an effective method of probing — it helps identify what, how, and why something happened. Let us consider an employee turnover rate in our organization is increasing. Why are Employees not satisfied? Why do Employees feel that they are underpaid? Why are Other employers paying higher salaries? Why Demand for such employees has increased in the market? Who are impacted by this problem? What will happen if this problem is not solved? What are the impacts? Where and When does this problem occur? Why is this problem occurring? How should the process work? How are people currently handling the problem? Fo rmulat ing t he hypo t hesis: Br eak do wn pr o blems and fo r mulate hypo theses. Fr ame the Questions which need to be answered or topics which need to be explored in order to solve a problem. Develop a comprehensive list of all possible issues related to the problem Reduce the comprehensive list by eliminating duplicates and combining overlapping issues Using consensus building, get down to a major issues list.

Data Collection: In order to answer the key questions and validate the hypotheses collection of realistic information is necessary. Depending on the type of problem being solved, different data collection techniques may be used. Data collection is a critical stage in problem solving - if it is superficial, biased or incomplete, data analysis will be difficult. Data Collection Techniques: Using data that has already been collected by others Systematically selecting and watching characteristics of people, objects or events. Oral questioning of respondents, either individually or as a group. Collecting data based on answers provided by respondents in written form. Facilitating free discussions on specific topics with selected group of participants.

Data Exploration: Before a formal data analysis can be conducted, the analyst must know how many cases are in the dataset, what variables are included, how many missing o bser vatio ns ther e ar e and what g ener al hypo theses the data is likely to suppo r t. An initial explo r atio n o f the dataset helps answer these questio ns by familiar izing analysts abo ut the data with which they are working. Analysts commonly use visualization for data exploration because it allows users to quickly and simply view most of the relevant features of their dataset. By doing this, users can identify variables that are likely to have interesting observations. By displaying data graphically through scatter plots or bar charts users can see if two or more variables correlate and determine if they are good candidates for further in-depth analysis.

Data Preparation: Data comes to you in a form that is not easy to analyze. We need to clean data and check it for consistency, extensive manipulation of the data is needed in order to analyze. Combining data Splitting data into many datasets. Then, identify the data type and category of the variables. Univariate Analysis: At this stage, we explore variables one by one. Method to perform Univar iate analysis will depend o n whether the var iable type is categ o r ical o r co ntinuo us. Continuous Variables: In the case of continuous variables, we need to understand the central tendency and spread of the variable. These are measured using diverse statistical metrics visualization methods. Categorical Variables: For categorical variables, we use a frequency table to understand the distribution of each category.

We can also read as a percentage of values under each category. It can be measured using two metrics, Count and Percent against each category. Model Building: This is really the entire process of building the solution and implementing the solution. The majority of the project time spent in the solution implementation stage. One interesting thing to remember with an analytical approach is that an analytical approach when you are building models, analytical models, is a very iterative process because ther e is no such thing as a final so lutio n o r a per fect so lutio n.

Typically, yo u will spend time building multiple models on multiple solutions before arriving at the best solution that the business will work with. There are many ways of taking decisions from a business perspective. Analytics is one way. There are others ways of taking a decision. It could be experience based decision taking. It could be gut-based decision making. And not every single time you will always choose an analytical approach. However, in the long run, it makes sense to build analytical capability because that leads to more objective decision making.

But fundamentally if you want to data to drive decision making, you need to make sure that you have invested in collecting the right data to enable your decision-making through data. Remember that an analytical problem-solving approach, which is different from the standard problem-solving approach. We need to remember these points: There is a clear confidence on data to drive solution identification.

We are using analytical techniques based on numeric theories. Yo u need to have a g o o d under standing o f theo r etical co ncepts to business situatio ns in order to build a feasible solution. What that means is you need to a good understanding of the business situation and the business context and as well a strong knowledge of analytical approaches and be able to merge the concepts, come up with a workable solution. In some industries, the rate of change is very high.

So, solutions age very fast. In other industries, the rate of change may not be as high and when you build a solution, you may have years where your solution works well but post that will need to be tweaked to manage the new business conditions. But, the way to assess whether or not your solution is working, is to periodically check solution effectiveness. You need to track dependability over time and you may need to make minor changes to bring the solution back on track. Sometimes, may have to build an entire solution from scratch because the environment has changed so dramatically that the solution that you built does not clutch anymore in the current business context. He may lack the knowledge and experience that you have. Since most problems are not unique, We may be able to corroborate the problem and possible solutions against other sources. The best solutions to a problem are often too difficult for the client to implement.

So be cautious about recommending the optimal solution to a problem. Most explanations require some degree of conciliation for execution. R is a simple programming language which includes many functions for Data Analytics, it has an effective data handling and storage facility. R provides graphical facilities for data analysis and reporting. I request you to please Install R and R studio which is freely downloadable.

Here in my book I use the code written in R studio. Scripts: Serves as an area to write and save R code 2. Workspace: Lists the datasets and variables in the R environment 3. Plots: Displays the plots generated by the R code 4. Console Provides a history of the executed R code and the output. R can per fo r m mathematical calculatio ns witho ut o blig atio n that yo u need to sto r e it in an object. The result is printed on the console. Anything written after sign will be considered as comments in R.

Type the following commands and understand the difference. Variables: We can store values into a variable to access it later. Try printing the current value of Y. If you wrote this code, congratulations! You wrote the first code in R and created an object. Functions: We can call a function by typing its name, followed by arguments to that function in parenthesis. Try the sum function, to add up a few numbers. Enter: sum 1, 3, 5 9 We use sqrt function to get the square root of Type the following commands and check the answers log 1 0 log 10 2. R data file using the save. An existing. R data file can be loaded using the load. Assume that we stored some sample scripts, We can list the files in the current directory from within R, by calling the list. For this first, we want to know what is the current directory R is using by default.

How to Make Big Data Analytics Incredibly Easy

By the end of this book, you will have learned how to implement machine learning algorithms and sell the results to your customers without writing a line of code. His research investigates the essential components of Big Data as a phenomenon and the impact of AI and Data Analytics on companies and people. He is the author of popular science books on data analytics and various research papers in international journals. About this book Data analytics has become a necessity in modern business, and skills such as data visualization, machine learning, and digital storytelling are now essential in every field.

Publication date: August Publisher Packt. Pages Milo D. Koretsky 0. Randall D. Knight 0. George Odian 0. John Kenkel 0. Trott 0. Carl S. Warren 2. Warren 0. Abraham Silberschatz 1. Frederick S. Hillier 1. William Stallings 0. Morris Mano 1. David Irwin 0. Morris Mano 0. Michael F. Ashby 0. William Thomson 0. Gene Mathers 0. Jack C. McCormac 1. William T. Segui 0. Richard T. Evans 0. Bill W. Tillery 0. Giorgio Rizzoni 0. Khurmi 1. Singiresu S. Rao 2.

Data Analytics Made Easy | Packt

He is the author of popular science books on data analytics and various research papers in international journals. About this book Data analytics has become a necessity in modern business, and skills such as data visualization, machine learning, and digital storytelling are now essential in every field. Publication date: August Publisher Packt. Pages ISBN Browse publications by this author. In this example for organic searches by date, we dragged over keywords to analyze traffic breakdown by keywords and fiscal quarter. If you want to do further analysis, you can drag and drop more data fields to the chart. If you want to analyze keyword breakdown by social network, for example, you can just drag over the Social Network field.

You have to create a new widget and input several variables just to create a bar chart. You need an easy way to get this data to the rest of your team, instead of it staying siloed away. When your data changes, you need everyone on your team to be updated with the latest results so they can act on it. We designed DataHero to end this problem of siloed data. It draws data from the applications your business already uses and brings it together into an all-in-one dashboard that you can customize and share. Your team can collaborate, your clients can access your dashboard, and everyone will see the latest results of your analysis. Your data is constantly changing, especially if you do several iterations of your keyword campaign, for example.

Your marketing team can see the latest results of your AdWords campaign and adjust their strategy accordingly. BIME is a business intelligence tool that also has the capability to create and share dashboards. You have to create a query and build widgets from your data. You spend more time customizing and resizing your dashboard widgets before you actually publish and share the latest results with your team. With a simple drag and drop interface, DataHero can give you more insights on how to segment your emails or build your keyword campaigns. You can easily integrate third-party apps like Dropbox or Marketo to gain further insights from your cloud-based data. But once you have those valuable insights, you can predict customer behavior and ultimately drive revenue. If you want to automatically update your charts, combine data, and create unlimited dashboards, click here to learn more about our Premium subscription.

Page composed with the free online HTML editor. Please subscribe for a license to remove these messages from the edited documents. Toggle navigation. The DataHero Blog. Intuitive navigation for every step of your analysis Data helps people make fast autonomous decisions.

Big Data Analytics Made Easy. Learn Python programming with free PDF books at ering

Michael F. Ashby 0. William Thomson 0. Gene Mathers 0. Jack C. McCormac 1. William T. Segui 0. Richard T. Evans 0. Bill W. Tillery 0. Giorgio Rizzoni 0. Khurmi 1. Singiresu S. Rao 2. Ron Larson 0. Ron Larson 1. Lakshmi Prasad — 1st Edition. Data Mining — Ian H. Witten, Frank Eibe — 2nd Edition. Boylestad — 9th Edition. Do you know that they are looking at risk or revenue? It really depends on the business interest; it depends on the business goals for that year.

In fact, a good customer this year may be a bad customer next year. This is why it is important to obtain the context or the problem statement before starting on an analysis. But this is not the only problem with this problem statement. Another problem is thinking about the decision which is, can you really insist on receiving g o o d applicatio ns o r can yo u insist o n appr o ving g o o d applicatio ns. Is the decisio n at the application stage or the approval stage? Can you really control applications to be good or can you control the decisions to enable only good customers to come on to you?.

Another problem with this problem statement is that we only want to receive credit card applications from good customers. Is it realistic for you to assume that you will have a solution that will never accept a bad customer? Again, not a realistic outcome. Coming back to our problem definition state which is, given a business problem, I want to get good customers as a credit card company. How do you frame that problem into something that analytical approach can tackle?

One way is to add specifics to the problem statement. So, think about specific, measurable, attainable, realistic, and timely outcomes that you can attach to that problem statement. That is, why we emphasize that you need to understand the business context thoroughly and talk to the business that you are tackling the right problem. How would I be able to add specifics to this pr o blem statement? Let us assume that I am lo o king at it fr o m the r isk per spective, because in this year my credit card companies focused on reducing the portfolio risk. So, I could have a variety of business problem statements. For example, reduce losses from cr edit car d default by at least 30 per cent in the fir st 12 mo nths po st implementatio n o f the new strategy. Develop an algorithm to screen applications that do not meet good customer defined criteria that will reduce defaults by 20 percent in the next 3 months. Identify strategies to reduce defaults by 20 percent in the next three months by allowing at-risk customers additional payment options.

We have decided that the good problem definition is something that we ar e tackling fr o m a r isk per spective. But, fo r the same business statement, we no w have three different problem statements that are tackling three different things. Again, which of these should I choose as a starting point for my analysis? Should I identify strategies for my existing customers or should I look at identifying potential new customers? Again, this is something that may be driven by business needs. So, it is important to constantly talk to the business to make sure that when you are starting an analytics project you are tackling the right problem statement. Getting to a clearly defined problem is often discovery driven — Start with a conceptual definition and through analysis root cause, impact analysis, etc.

A problem becomes known when a person observes a discrepancy between the way things are and the way things ought to be. Root Cause Analysis is an effective method of probing — it helps identify what, how, and why something happened. Let us consider an employee turnover rate in our organization is increasing. Why are Employees not satisfied? Why do Employees feel that they are underpaid? Why are Other employers paying higher salaries? Why Demand for such employees has increased in the market? Who are impacted by this problem? What will happen if this problem is not solved? What are the impacts? Where and When does this problem occur? Why is this problem occurring? How should the process work? How are people currently handling the problem? Fo rmulat ing t he hypo t hesis: Br eak do wn pr o blems and fo r mulate hypo theses.

Fr ame the Questions which need to be answered or topics which need to be explored in order to solve a problem. Develop a comprehensive list of all possible issues related to the problem Reduce the comprehensive list by eliminating duplicates and combining overlapping issues Using consensus building, get down to a major issues list. Data Collection: In order to answer the key questions and validate the hypotheses collection of realistic information is necessary. Depending on the type of problem being solved, different data collection techniques may be used. Data collection is a critical stage in problem solving - if it is superficial, biased or incomplete, data analysis will be difficult.

Data Collection Techniques: Using data that has already been collected by others Systematically selecting and watching characteristics of people, objects or events. Oral questioning of respondents, either individually or as a group. Collecting data based on answers provided by respondents in written form. Facilitating free discussions on specific topics with selected group of participants. Data Exploration: Before a formal data analysis can be conducted, the analyst must know how many cases are in the dataset, what variables are included, how many missing o bser vatio ns ther e ar e and what g ener al hypo theses the data is likely to suppo r t. An initial explo r atio n o f the dataset helps answer these questio ns by familiar izing analysts abo ut the data with which they are working. Analysts commonly use visualization for data exploration because it allows users to quickly and simply view most of the relevant features of their dataset.

By doing this, users can identify variables that are likely to have interesting observations. By displaying data graphically through scatter plots or bar charts users can see if two or more variables correlate and determine if they are good candidates for further in-depth analysis. Data Preparation: Data comes to you in a form that is not easy to analyze. We need to clean data and check it for consistency, extensive manipulation of the data is needed in order to analyze. Combining data Splitting data into many datasets. Then, identify the data type and category of the variables. Univariate Analysis: At this stage, we explore variables one by one.

Method to perform Univar iate analysis will depend o n whether the var iable type is categ o r ical o r co ntinuo us. Continuous Variables: In the case of continuous variables, we need to understand the central tendency and spread of the variable. These are measured using diverse statistical metrics visualization methods. Categorical Variables: For categorical variables, we use a frequency table to understand the distribution of each category. We can also read as a percentage of values under each category. It can be measured using two metrics, Count and Percent against each category. Model Building: This is really the entire process of building the solution and implementing the solution. The majority of the project time spent in the solution implementation stage. One interesting thing to remember with an analytical approach is that an analytical approach when you are building models, analytical models, is a very iterative process because ther e is no such thing as a final so lutio n o r a per fect so lutio n.

Typically, yo u will spend time building multiple models on multiple solutions before arriving at the best solution that the business will work with. There are many ways of taking decisions from a business perspective. Analytics is one way. There are others ways of taking a decision. It could be experience based decision taking. It could be gut-based decision making. And not every single time you will always choose an analytical approach.

However, in the long run, it makes sense to build analytical capability because that leads to more objective decision making. But fundamentally if you want to data to drive decision making, you need to make sure that you have invested in collecting the right data to enable your decision-making through data. Remember that an analytical problem-solving approach, which is different from the standard problem-solving approach. We need to remember these points: There is a clear confidence on data to drive solution identification. We are using analytical techniques based on numeric theories. Yo u need to have a g o o d under standing o f theo r etical co ncepts to business situatio ns in order to build a feasible solution. What that means is you need to a good understanding of the business situation and the business context and as well a strong knowledge of analytical approaches and be able to merge the concepts, come up with a workable solution.

In some industries, the rate of change is very high. So, solutions age very fast. In other industries, the rate of change may not be as high and when you build a solution, you may have years where your solution works well but post that will need to be tweaked to manage the new business conditions. But, the way to assess whether or not your solution is working, is to periodically check solution effectiveness. You need to track dependability over time and you may need to make minor changes to bring the solution back on track. Sometimes, may have to build an entire solution from scratch because the environment has changed so dramatically that the solution that you built does not clutch anymore in the current business context.

He may lack the knowledge and experience that you have. Since most problems are not unique, We may be able to corroborate the problem and possible solutions against other sources. The best solutions to a problem are often too difficult for the client to implement. So be cautious about recommending the optimal solution to a problem. Most explanations require some degree of conciliation for execution. R is a simple programming language which includes many functions for Data Analytics, it has an effective data handling and storage facility. R provides graphical facilities for data analysis and reporting.

I request you to please Install R and R studio which is freely downloadable. Here in my book I use the code written in R studio. Scripts: Serves as an area to write and save R code 2. Workspace: Lists the datasets and variables in the R environment 3. Plots: Displays the plots generated by the R code 4. Console Provides a history of the executed R code and the output. R can per fo r m mathematical calculatio ns witho ut o blig atio n that yo u need to sto r e it in an object. The result is printed on the console. Anything written after sign will be considered as comments in R. Type the following commands and understand the difference. Variables: We can store values into a variable to access it later. Try printing the current value of Y. If you wrote this code, congratulations!

You wrote the first code in R and created an object. Functions: We can call a function by typing its name, followed by arguments to that function in parenthesis. Try the sum function, to add up a few numbers. Enter: sum 1, 3, 5 9 We use sqrt function to get the square root of Type the following commands and check the answers log 1 0 log 10 2. R data file using the save. An existing. R data file can be loaded using the load. Assume that we stored some sample scripts, We can list the files in the current directory from within R, by calling the list.

For this first, we want to know what is the current directory R is using by default. We set the folder R data in D drive as working directory. R provides several kinds of data structure each designed to optimize some aspect of storage, access, or processing. Examples of Data structures Vector 2. Matrix 3. Factor 4. Data Frame 1. Vectors Vectors are a basic building block for data in R. R variables are actually vectors. A vector can only consist of values in the same class. The tests for vectors can be conducted using the is. The name may sound frightening, but a vector is simply a list of values. R provides functionality that enables the easy creation and manipulation of vectors. The following R code illustrates how a vector can be created using the combine function, c or the colon operator, :, Let us create a vector of numbers: c 4,7,9 The c function c is short for Combine creates a new vector by combining a list of values.

Let us build a vector from the sequence of integers from 5 to 9. Try getting the first and fourth words: sentence [c 1, 4] This means you can retrieve ranges of values. Get the second through fourth words: sentence [] We can set ranges of values, by just providing the values in a vector. Matrices A matrix in R is a collection of homogeneous elements arranged in 2 dimensions. A matrix is a vector with a dim attribute, i. Rows and columns can have names, dimnames , rownames , colnames Let us look at the basics of working with matrices, creating them, accessing them and plotting them. Let us create a matrix 3 rows high by 4 columns wide, with all its fields set to 0.

Look at the following code, the content of Sample is filled with the columns consecutively. Try retrieving the Third row: Sample [3,] [1] 3 7 11 15 To get an entire column, omit the row index. Retrieve the fourth column: Sample [,4] [1] 13 14 15 16 3. Factors When we want the data to be

grouped by category, R has a special type called a factor to track these categorized values. A factor is a vector whose elements can take on one of a specific set of values. The set of values that the elements of a factor can take is called its levels. Pass the factor to the as. Data Frames data frames provide a structure for storing and accessing several variables of possibly different data types. Because of their flexibility to handle many data types, data frames are the preferred input format for many of the modeling functions available in R.

Let us create three individual objects named Id, Gender, and Age and tie them together into a data set. It has a specific number of columns, each of which is expected to contain values of a particular type. It also has an indeterminate number of rows - sets of related values for each column. We can get individual columns by providing their index number in double-brackets. Impo rt ing T XT files: If yo u have a. Call read. The first line is not automatically treated as co lumn header s with r ead. This behavio r is co ntr o lled by the header ar g ument.

CSV function. The first argument specifies which data frame in R is to be exported. The next argument specifies the file to be created. Since we do not wish to include row names we were given option row. As we have shown in the example it is very common not to want the quotes when creating a text file. Data exploration will help you to create accurate models if you perform this in a planned way. Before a formal data analysis can be conducted, the analyst must know how many cases are in the dataset, what variables are included, how many missing observations there are in the dataset.

Data exploration Steps includes Understanding the datasets and variables, Checking attributes of the data, Recognize and treat missing values, o utlier s, Under standing basic pr esentatio n o f the data etc. Data explo r atio n activities include the study of the data in terms of basic statistical measures and creation of graphs and plots to visualize and identify relationships and patterns. An initial explo r atio n o f the dataset helps answer these questio ns by familiar izing analysts abo ut the data with which they ar e wo r king.

Additio nal questio ns and co nsider atio ns fo r the data co nditio ning step includes, What ar e the data so ur ces? What ar e the tar g et fields? Ho w clean the data is? How consistent are the contents and files? Being a Data Scientist, you need to determine to what degree the data contains missing or inconsistent values and if the data contains values deviating from normal and Assess the consistency of the data types. For instance, if the team expects certain data to be numeric, confirm it is numeric or if it is a mixture of alphanumeric strings and text. Review the content of data columns or other inputs, and check to ensur e they make sense. Fo r instance, if the pr o ject invo lves analyzing inco me levels, preview the data to confirm that the income values are positive or if it is acceptable to have zeros or negative values.

Look for any evidence of systematic error. Examples include data feeds from sensors or other data sources breaking without anyone noticing, which causes invalid, incorrect, or missing data values. Review the data to gauge if the definition of the data is the same for all measurements. In some cases, a data column is repurposed, or the column stops being populated, without this change being annotated or without others being notified. After the team has collected and obtained at least some of the datasets needed for the subsequent analysis, a useful step is to lever ag e data visualizatio n to o ls to lo o k at hig h-level patter ns in the data enables one to understand characteristics about the data very quickly.

One example is using data visualization to examine data quality, such as whether the data contains many unexpected values o r o ther indicato r s o f dir ty data. Ano ther example is Skewness, such as if the majority of the data is heavily shifted toward one value or end of a continuum. Data Visualizatio n enables the user to find ar eas o f inter est, zo o m, and filter to find mo r e detailed information about a particular area of the data, and then find the detailed data behind a particular area. This approach provides a high-level view of the data and a great deal of information about a given dataset in a relatively short period of time. Fo r instance, did custo mer lifetime value chang e at so me po int in the middle o f data co llectio n?

Or if wo r king with financials, did the inter est calculatio n chang e from simple to compound at the end of the year? Does the data distribution stay consistent over all the data? Data Analytics Made Easy is an accessible guide to help you start analyzing data and quickly apply these skills to your work. The book introduces the concepts of data analytics and shows you how to get your data ready and apply ML algorithms.

Create impressive visualizations with Microsoft Power BI and learn the greatest secret in successful analytics — how to tell a story with your data. By the end of this book, you will have learned how to implement machine learning algorithms and sell the results to your customers without writing a line of code.