BIANCA DATA

DATA SET STARTER KIT

MEET MY FRIEND “RIC” & LET’S START...

I’m about to introduce you to my friend who helps me with every project I do. His name is RIC (pronounced Rick), and he’s a really cool guy ;)

WRITTEN BY

BIANCA JORDAN

Starting on a Data Set...

The RIC Reminder:

R - RETRIEVE & REFER

I - IDENTIFY & INTUIT

C - CLEAN & CONCENTRATE

“

RIC will be your guide every time you start on a new project. He’s not just going to show you the practical steps when you need to import and clean your data, he’s going to set you up to better understand your project.

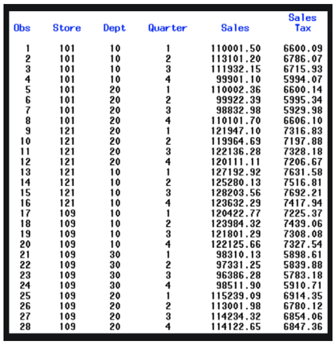
Oh, here he is, now. How rude of me, RIC, meet my friend, my friend, meet RIC...

R

Retrieve & Refer

Hello!

My name is RIC (pronounced “Rick”) and I’m here to help you wrap your mind around a new data set. I know it can be intimidating when someone gives you a bunch of numbers and expects you to solve all their problems, but let’s see if my little system can help you organize both your data and your thoughts.



```
                                                         Sales
Obs    Store    Dept    Quarter        Sales             Tax

  1     101      10        1         110001.50       6600.09
  2     101      10        2         113101.20       6786.07
  3     101      10        3         111932.15       6715.93
  4     101      10        4          99901.10       5994.07
  5     101      20        1         110002.36       6600.14
  6     101      20        2          99922.39       5995.34
  7     101      20        3          98832.98       5929.98
  8     101      20        4         110101.70       6606.10
  9     121      20        1         121947.10       7316.83
 10     121      20        2         119964.69       7197.88
 11     121      20        3         122136.28       7328.18
 12     121      20        4         120111.11       7206.67
 13     121      10        1         127192.92       7631.58
 14     121      10        2         125280.13       7516.81
 15     121      10        3         128203.56       7692.21
 16     121      10        4         123632.29       7417.94
 17     109      10        1         120422.77       7225.37
 18     109      10        2         123984.32       7439.06
 19     109      10        3         121801.29       7308.08
 20     109      10        4         122125.66       7327.54
 21     109      30        1          98310.13       5898.61
 22     109      30        2          97331.25       5839.88
 23     109      30        3          96386.28       5783.18
 24     109      30        4          98511.90       5910.71
 25     109      20        1         115239.09       6914.35
 26     109      20        2         113001.98       6780.12
 27     109      20        3         114234.32       6854.06
 28     109      20        4         114122.65       6847.36
```

RIC stands for Retrieve & Refer, Identify & Intuit, Clean & Concentrate – people call me RIC because it’s easy to remember – but let’s dive into the details below. You’ll see that each letter of my name stands for a simple step in your project, but some pretty big ideas too…

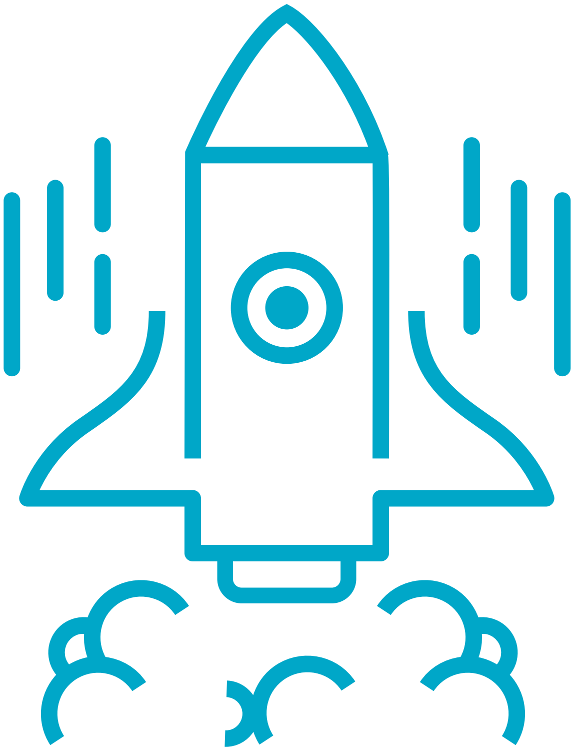
RETRIEVE THE DATA.

Some practical things first: make sure you data is in the right format (the easiest data formats to work with are .csv and .txt).

Then import into your choice of IDE, I usually use Spyder which is part of Anaconda, because it has the console incorporate and it is easier to check your progress as you run through your code.

Another good choice for IDE is JupyterNotebook as it also has a user friendly interface. Once you get more comfortable coding, any text editor will do as long as you run your code in your command line.

And so in order to be able to read your data import it using the Pandas library... *hint* make a new variable, call it whatever you want and follow the line pd.read_csv (for csv and txt files)



REFER TO THE GOAL.

Alright, so you have the data where you want it... I know the temptation is to just start organizing and cleaning, but take a beat: can you easily define your goal? I like to refer to the original question the client brought me, for instance: Can we predict our customer's lifetime value? Write down the goal as simply as you can so you can keep referring back to it.

Identify & Intuit

IDENTIFY THE IMPORTANT VARIABLES.

Take a second to understand your variables before digging into the numbers. You aren't a computer, you have intuition and you should use it: do you already recognize some important variables to keep an eye on? Do you have some understanding of the company or the industry or the community this data set is based on?

INTUIT POSSIBLE EXPLANATIONS.

I believe human intuition is underrated in data science. The general consensus is that intuition gets you in trouble because you smuggle your biases into the data. That's absolutely true and let's be sure to not jump to conclusions. BUT, all intuition means is "recognition," and humans are pretty great at intuitively creating explanations, so let's lean into it. Let's recognize that we might have certain expectations within the data, and let's be as honest as we can about whether those expectations are met or subverted.

In practical terms, if we refer to the example goal of predicting a customer's lifetime value, our intuitions can already start telling us a story: maybe variables like age and annual income might be more important than weight or height. You don't have to ignore those intuitions.

Again, this isn't about jumping to conclusions, it's about keeping your background thoughts organized as you dive into the data set in order to let it tell you its story, which brings us to the next step…

C

Clean & Concentrate

CLEAN THE DATA.

Cleaning the data means to remove duplicates (look at your data and see what variables you do not need), remove or replace NaN values, change letters to lower – upper case, convert categorical (strings) values to numerical values, get rid of extra spaces.

PRO TIP Please please please keep a record of all the changes you make along the way. You ever go on a cleaning spree at home and then forget where you put your car keys? Keep a record of the changes because no one cares about how clean the house is when they can't make sense of where or what anything is.

CONCENTRATE YOUR FOCUS.

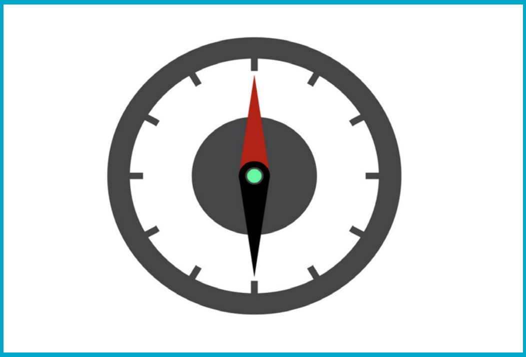
It's no secret that cleaning is the most tedious part of the process. But that's why I want you to be thinking about the broader story of the data in the background. Which variables are the most important for getting complete data for? Can we remove some columns of variables completely and concentrate the crucial ones? Narrow down your data by extracting date components, joining data sets – manipulate your data but be careful not to import unintended bias).

CHEERS!

Not gonna lie, now ol' Ric is ready for a drink... But don't forget...

THIS IS JUST THE PRE-PARTY...

Alright, feeling clean and concentrated? I know that seems like just a lot of prep work for the big party to come – namely visualizing and predicting with your data – but don't underestimate the importance of this prep work. Think of RIC like a compass. Every time you get lost in all the variables and numbers of your data set, take a look at your friend RIC, and hopefully we can help you find North on your project.

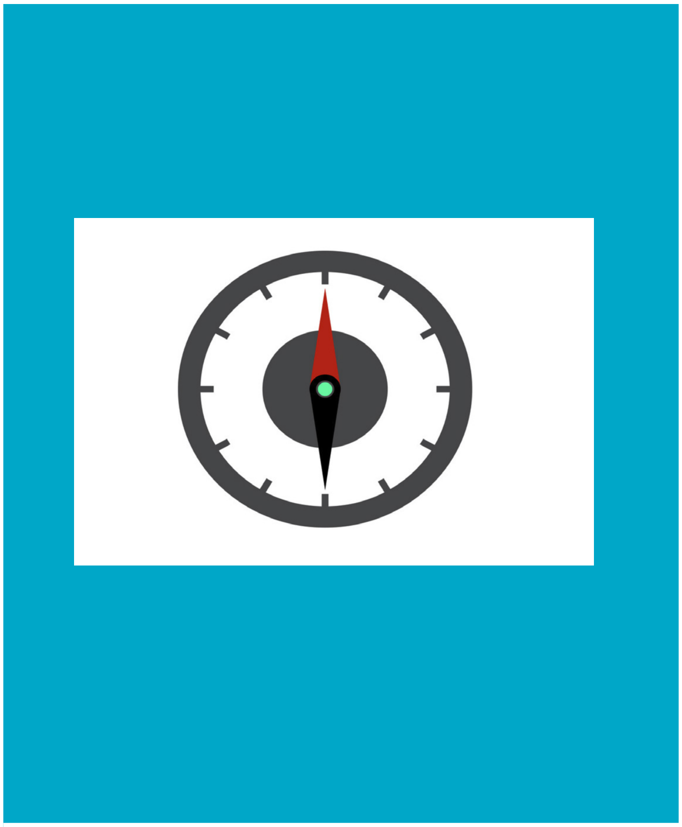


*Think of RIC like a compass for whenever you're feeling a bit lost in your project.

NEXT STEPS:

I hope you loved meeting RIC and you have a better understanding of your data set. But if you're still a little confused on what to do next, don't worry, I got your back. As a subscriber at BiancaData.com you're going to be the first to know when I release my Data Science for Beginners program and I'll take you through every step of this process in an unforgettable way. Stay tuned, and in the meantime if you have any questions, send me an email at info@biancadata.com and thank you so much for being a part of this supportive community!

BiancaData.com



Thanks for making me a part of your Data Science journey :)