dasca™
DATA SCIENCE COUNCIL OF AMERICA

21 POWERFUL TIPS, TRICKS, AND HACKS FOR DATA SCIENTISTS

DATA SCIENCE COUNCIL OF AMERICA

CONTENTS

DATA SCIENCE COUNCIL OF AMERICA

INTRODUCTION

Data is the oil of the 21st century, and people who can analyze it are the high-performing combustion engines who extract it, refine it, and use it effectively to make it valuable.

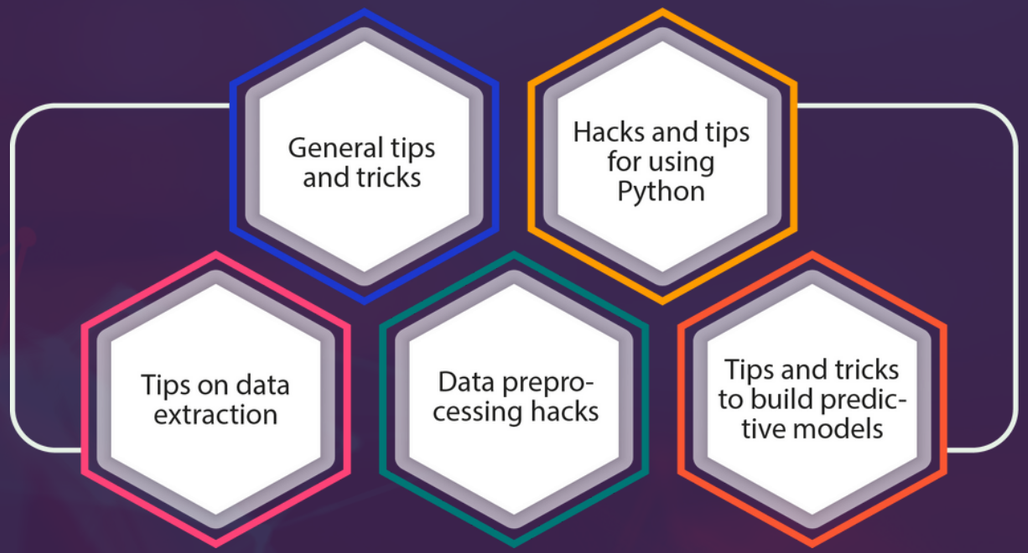
Jobs in data science are some of the best-suited vocations to thrive in this era. The most appealing aspect of a career in data science is it opens the world to you. Data is the new age gatekeeper that can get you into about any arena.

From Amazon to Facebook and down to the thousands of disruptors across industries, data science has become intrinsic to business success. Companies are culling through their troves of raw information to gain a competitive advantage. All are seeking a firm footing in data and analytics.

So how do these data wizards mine the terabytes of information out there?

This guide elucidates some time-saving hacks, tips, and tricks that data scientists can use to make their lives easier. It will help you make smarter business choices and become a better and more efficient data scientist.

Data Science is a growing discipline. There is still a lot to learn for both young and seasoned professionals. Reflecting on the decades of experience of Senior Data Scientists and industry best practices, this guide brings to you:



Being a data scientist is not only about data crunching. It's about understanding the business challenge, creating some valuable actionable insights to the data, and communicating their findings to the business.

JEAN-PAUL ISSON

VP-Chief Data Science & Artificial Intelligence Officer, SITA

A DATA SCIENTIST

They are the unicorns who bring a creative combination of an engineer, visualization expert, an analyst, a researcher, and a navigator between business and data science teams.

Data Scientists come from a diverse range of backgrounds – from engineering and mathematics to business and even social sciences. With a confluence of quantitative and analytical skills, they play a central role in providing strategic guidance with business intelligence. It's one of those job roles that may sound super-technical, mysterious, and even hard to get, but with the right push in the right direction, one can become a good data scientist.

Javis Miller, a data scientist at Spotify shares that data scientists worth their salt must be able to explain their contribution to business without using jargon.

"You can't simply go to a business stakeholder and say that I ran a logistic regression on this data to classify…," he says. On a related note, Heather Nolis, a machine learning engineer at T-Mobile says that for a better working relationship with engineers, data scientists must appreciate and try to speak the language of engineers.

They are, indeed, the bridge between technical and business; not only by function but by training.

Knowledge and *awareness* are the greatest strengths of a good data scientist. No matter how much you can code or how good you are at statistics – your business awareness and familiarity with uncertainty are the most valuable skills you can offer as a data scientist.

DATA SCIENCE COUNCIL OF AMERICA

DATA SCIENTISTS EMPLOYERS SEEK

In-demand Soft-Skills

- Business acumen
- Collaboration
- Communication
- Attention to detail
- Storytelling
- Critical thinking
- Leadership
- Problem Solving
- Project Management

In-demand Hard-Skills

- Statistics
- Data Modelling
- Data Visualization
- Mathematics
- Machine Learning
- R
- Python
- Natural Language Processing
- SQL

DATA SCIENCE COUNCIL OF AMERICA

21 TIPS, TRICKS, AND HACKS FOR DATA SCIENTISTS

This section is divided into five parts, representing the major phases of the data science process. Part one covers a few general tips important for data scientists, followed by Python, data extraction, and data preparation, and finally predictive modeling.

General

Data science has found its way into all industries and disciplines. Find some basic, yet critical tips and tricks to understand the strategies and practices followed by experts in data science. These tips will help you focus on what's necessary and discard the unnecessary.

Python

Python is the go-to language in data science and machine learning. Learn some 'pythonic' hacks to code better and easier. Given the universality of the language among data scientists, all the examples in this guide follow Python coding representation.

Data Extraction

Extracting data is the first step to bring your hypothesis to reality. Your ability to tell from where to aggregate the data and the skills to review it to eliminate the errors can make or break your model. Learn some tips and tricks to extract and clean the data to make it analysis-ready.

Data Pre-processing

Data preprocessing is an important step before model building. Normalizing the data, standardization, categorizing variables are just a few tasks performed at this stage. Learn a few hacks on the functions to use for normalization and testing if the data you have is enough.

Model Building

Building models is an important stage for data scientists. It guides the process of actually architecting the data for the business context. In these model building hacks, tips, and tricks explore the libraries you can use for robust model testing and data visualization.

1. ASK THE RIGHT QUESTIONS TO SET GOALS

A data set will tell you no more than what you ask of it. Nurture and improve your knack for asking the right questions that lead to informative answers.

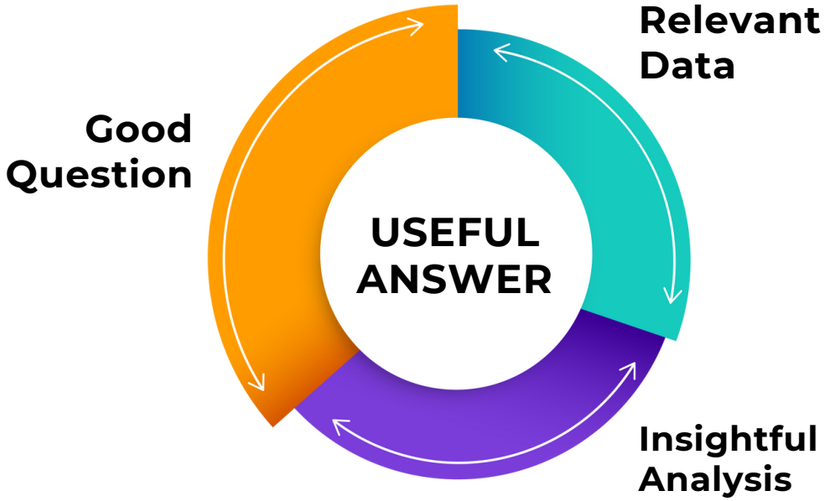
One way is to begin by investigating what do the customers expect. It may be related to a much-needed product improvement or a problem. Often customer expectations can lead to the questions and the answers that you must seek from the data.

A word of caution as you ask your questions: always acknowledge your assumptions and make sure they are correct. Your results will be as good as your assumptions. Recognizing your assumptions can be a challenging feat as often they are innate to our knowledge. Accomplish this by breaking the reasoning between your observation and conclusion into logical steps. For instance:

Observation: A has gone down recently.

Assumption: Because A always course-correct itself to a certain mean value, X,

Conclusion: A will soon go up toward X



2. SET TIME ASIDE FOR EXPLORATION OF NEW LIBRARIES AND FEATURES

Life can be busy as you move into a job. As a data scientist, make sure you intently take out time for leisure discoveries of new tools and updates that may not be directly related to your regular job. Going out of your comfort zone will unleash your imagination, keep your knowledge up-to-date, and can help you prepare yourself for uncertainties.

You could explore new libraries and find some incredible tools that save you time. **Pandas Profiling,** for instance, can be a great package when working with smaller datasets. It conducts exploratory data analysis and converts it into a report. All in one line of code! **Gradio** is another package that helps you develop and deploy a web app for ML model in just three lines of code. These are new developments. Unless you take time out to discover the new releases and integrate yourself into data science community, it will be difficult to find these gems.

Top 10 Data Science Tools You Should Know:

3. FOLLOW THE BEST CODING PRACTICES

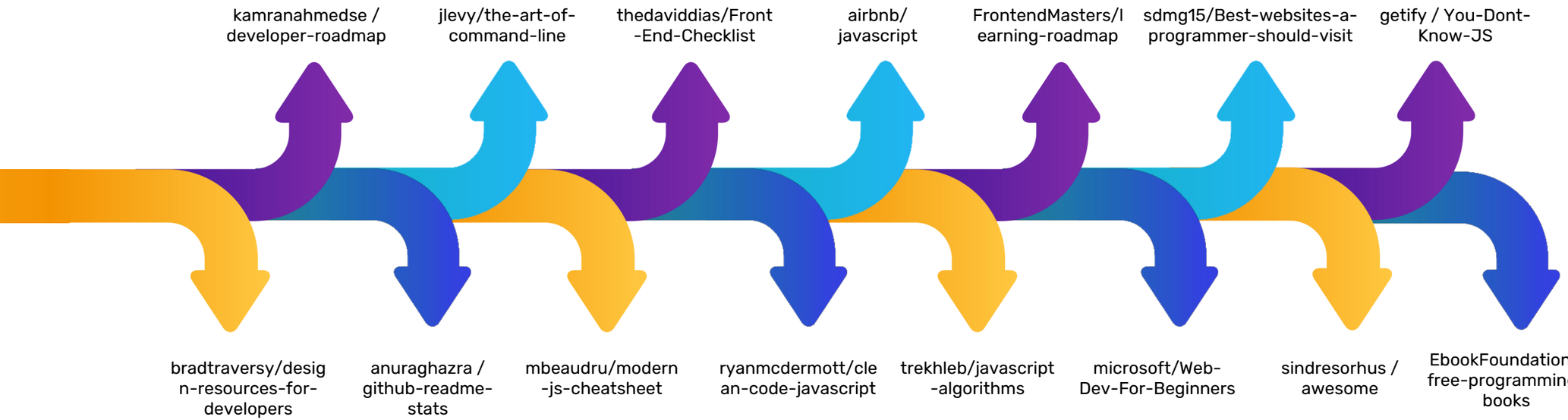
It’s always advisable to follow the best practices for code documentation and preservation for your and others’ sake lest someone else has to work on your project and code in the future. Make sure everyone in your team abide by these rules. Here are a few best practices you must follow.

- Use proper documentation to write a short description of what the code and the project are about and how to use it.
- Add comments in between your code to make it legible for fellow programmers.
- Make sure your variables, files, folders, and documents have meaningful names. (It’s one of the best ways to stay organized.)
- Consider naming a scheme and sticking to it throughout your code.
- Try to import all libraries at the beginning of your notebook. (You can add a comment on why you are loading each.)
- Maintain proper spacing between your code lines.
- If your code is becoming too heavy, you can hide a part of it that’s not important right now. It will make your notebook cleaner.
- Use versioning and code repositories. Modern repos are based on versioning systems, which track your code for the changes you make.

4. USE GIT AND GITHUB

From learning version control to building your data science portfolio, Git and GitHub are the important resources you must develop a good working knowledge of. Github.com can provide you free web hosting of code repos in a version control system. It also facilitates collaboration among programmers. You can decide to keep it public or private. You can find online tutorials on how to get started. Many seasoned data scientists also use GitHub repositories to demonstrate their capabilities to do a job by showcasing their contribution through GitHub.

TOP 15 MOST VALUABLE GITHUB REPOS IN 2021



5. KNOW THE MOST IMPORTANT LANGUAGES – R, PYTHON, SQL, AND MATLAB

R, Python, SQL, and MATLAB are a must-know for every programmer: data analyst or data scientist. A programming language enables you to design programs specific to your needs. You can then reuse these programs for projects whenever you require taking the same action.

R and Python are two of the most important programming languages used for all purposes. To begin with, you can choose either of them to develop your programming skills. Their biggest advantages are:

- They can manipulate data and work with multiple data types and software, and
- They can give end-to-end solutions for business problems.

One of the limitations of R and Python, though, is they cannot work with specific domains such as relational database management systems (RDBMS). **SQL** is used for that purpose. It can be used for querying from RDBMS and also in preparation for BI analysis. This makes SQL a must-know language for data scientists.

Another language that is essential to data science is **MATLAB.** It is used for working with matrices and tables. However, it is a paid service, which is its biggest limitation. Either way, these four languages cover most of the languages and tools used when working with data.

That said, when dealing with Big Data, apart from R and Python, faster languages like **Java** and **Scala** are often used. The latter two are very useful when combining data from a variety of sources.

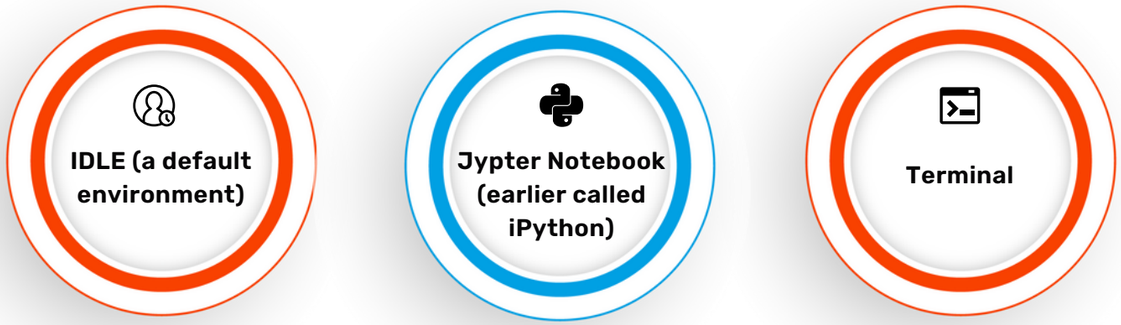
Some important software: Software is usually created by a combination of these languages and provides easy to use functionalities (check Pro Tip on the sidebar).

MS Excel and SPSS are some examples of software used in analytics. For working with Big Data, applications such as Apache Hadoop and MongoDB are used. Power BI, Tableau, SAS, and Qlik are some of the top-notch BI visualization software.

6. USE IPYTHON OR JUPYTER NOTEBOOK FOR IDE

Once you install Python, you will need to pick a virtual or an integrated development environment (IDE). A virtual environment is one of the most important tools when working with Python. It keeps the dependencies of different projects separate by creating isolated virtual environments.

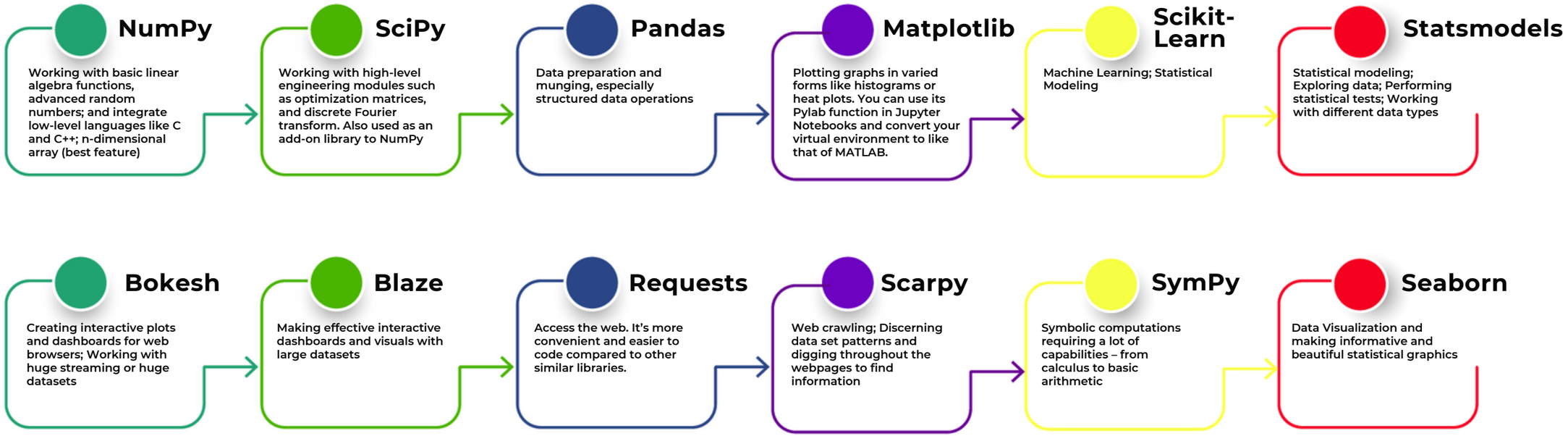
There are three options you can choose from:



We highly recommend that every data scientist use Jupyter Notebooks. From codes to presentations and generating full-fledged reports, Jupyter Notebooks provide a range of amazing features. Say, you write your code, you can run it in blocks than going line by line. It can also be used to perform data visualization. To create a presentation, you can create slides with super flexibility. To convert your notebook into slides, all you need is to choose Slideshow from Cell Toolbar. It has a powerful, versatile, and shareable interface. Jupyter IDE truly helps data scientists do it all.

7. KNOW THE 12 PYTHON LIBRARIES YOU WILL NEED

Python offers a gamut of useful libraries. Here are the libraries that data scientists use the most for analysis and computations. A few of them are also used in the examples used in the rest of this guide.



8. CREATE A ONE-LINE FUNCTION WITHOUT DEF KEYWORD (LAMBDA)

There may be times when you need to use a simple function for one time in the larger Python program. In such cases, going through the process of defining a function using def keyword, determining its name, and so on may look like a hassle.

If you are sure that you won't need to use this function after a certain program, you can use a shorter syntax that will allow you to focus on the function instead of other formalities associated with def keyword function. You can do it by using the anonymous function, also referred to as Lambda. Here's how the both differs.

A DEF KEYWORD FUNCTION:

```
def raise_to_the_power_of_3(x);
return x ** 3
```

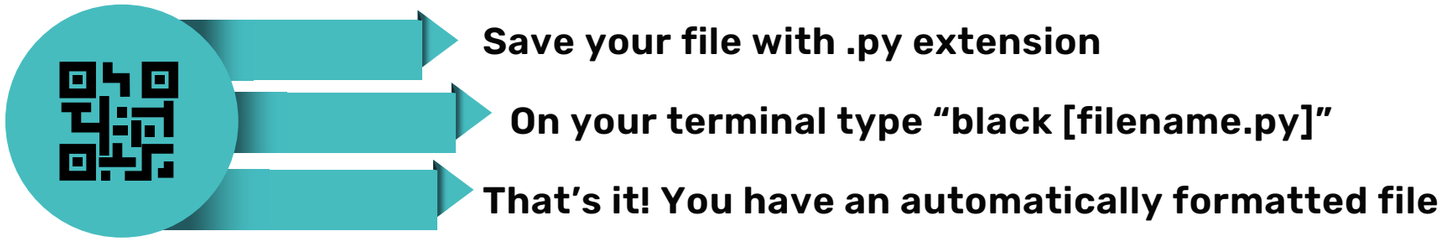
A LAMBDA FUNCTION:

```
raise_to_the_power_of_3 = lambda x: x ** 3
```

9. FORMAT YOUR CODE USING BLACK

Black is the uncompromising code formatter! If you are developing a lot of code, it can make your life so much easier by cleaning your code and making it easy to read for you and others. It also makes code reviewing faster.

Black is an automatic formatter for Python. All you have to do is write the code and then let Black format it. This is a great way to focus on your code instead of your structure. It's like waking up to a house that magically cleans itself. To invoke the power of Black code formatter:



Save your file with .py extension

On your terminal type "black [filename.py]"

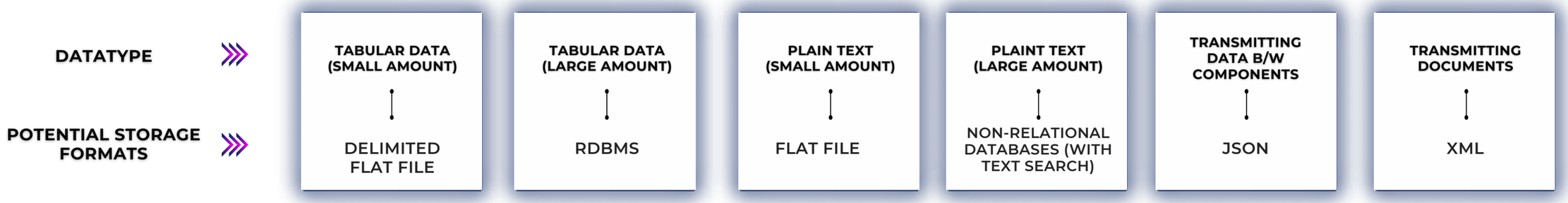
That's it! You have an automatically formatted file

10. KNOW YOUR DATA SOURCES

Albeit the playground of data engineers and analysts, data scientists are always concerned with scouting for data. Understanding the type of data available and how to store them is fundamental to data science. As a data scientist, your concerns would be: Where is the data now? In what shape is it? And How can I access it?

Data can be broadly divided into three forms. **Structured,** which includes transactional data as on relational database management systems. Its ready-made structure makes it easiest to work with. **Semi-structured** including text files, XML files, HTML files, among others that requires conversions into forms to makes it easier for analysis. Finally, **Unstructured data,** which includes audio, video feeds, web pages, and sensor data. It is the hardest to convert.

Additionally, there are three ways in which data scientists can access the data: **File, Database,** or **API.** In all three cases data may be stored in any format: flat files, XML, JSON, HTML, etc. Each type has its properties and qualities. Google search, Company websites, Kaggle, and such sources can help you scout for the data you need. In web searches, keep it in mind to use keywords such as ‘data’ and ‘API’. It can make a substantial difference.



Great data scientists never assume they know something without in-depth analysis, they think in hypotheses which need to be either rejected or proved, and they ask a lot of questions, even if they are 99.9% sure they know the answer.

KAROLIS URBONAS

Global Head of Machine Learning & Science, Amazon Web Services (AWS)

11. DON’T UNDERESTIMATE SQL

Often data science professionals underrate the importance of learning SQL. Data Scientists often work with databases, and fetching the data can become tiresome without a knowledge of SQL. It is the most prominent programming language to manipulate data and extract it. What more? It can be of value to a variety of data professionals – from data engineers and analysts to business analysts and data scientists.

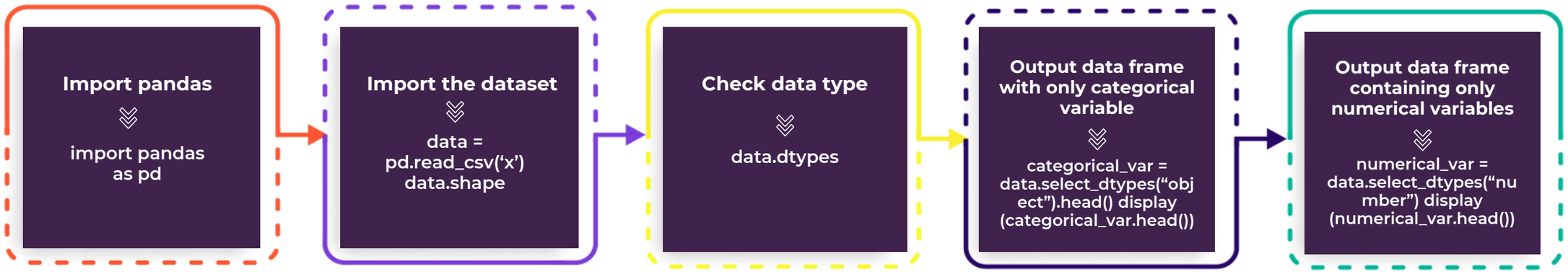
To give a crude example, say you want to analyze millions of customer orders in your company and forecast how the orders per day will change in the time to come. To do that, first, you will need to write a SQL query to fetch the orders placed each day. Then will come the use of R or Python to run a statistical forecast. It can be difficult to reach far in data analysis without knowing SQL. Along with RDBMS, SQL also works for new-age database systems such as MySQL, Oracle, Microsoft Access, IBM, among others.

SQL can be used to:

12. EXTRACT BY DATATYPE INTO DIFFERENT DATA FRAMES

Data Frames are a two-dimensional tabular representation of data, just like a spreadsheet with columns and rows. ‘Column names’ refer to the columns, and rows can be accessed using numbers 0 to n. Your data will be first read into a data frame, and then will become ready for other operations. Data frames are fundamental data models for pandas.

Extracting categorical data (object like name or ID) from numerical data (float and integers like income) into different data frames can be a useful feature at the time of data visualization and analysis. Here is a simple representation of how you can segregate your data by its type and represent it into different data frames in Python using Pandas.



13. REMOVE EMOJIS FROM TEXT

Once you extract your data, the first step is to clean the data to make it more understandable. The first step toward that is to remove unwanted characters such as emojis, links, unnecessary spaces, punctuations, among others. Given below is an easy way to remove emojis from your text data.

```
text= "Hi 😊! Is the group meeting tomorrow? I need to plan something.
So, please let me know.😊🙏"
processed_text=text.encode('ascii', 'ignore').decode('ascii')
#Display text with emoji
print ("Raw message:", text)
#Display text without emoji
print ("Processed message:", processed_text)
```

14. TEST YOUR DATA TO KNOW IF IT'S ENOUGH

Say you have the data; now how do you know if that's enough? Must you keep looking for more or should you attack your data and start with the processing and analysis?

To figure out the answer to this dilemma take a deeper look at your data. Often data sets aren't what they seem or what you imagine them to be.

For instance, say you have a goal of recommending perfumes to users of a perfume website based on the ratings the other users have provided for the perfumes. You begin with the question/hypothesis: do users like perfumes of certain scent types significantly more than others?

For this example, say you get access to a data set from a popular perfume-rating website wherein thousands of users have given ratings of one to five stars. Now, as you get to test your hypothesis with this data, you realize that the CSV file contains only three columns: USER_NAME, PERFUME_NAME, and RATING. (Drat! No scent types.) The dataset that may have seemed very relevant would now seem less so. Thus, to find the answer, you must find either a data set matching perfume names to scent types or try to infer it from the data you have by searching manually from each perfume name.

To find out if your data is enough, before spending much time manipulating your data or diving into analysis, perform a spot check. To do this, pick a few data points and attempt to arrive at the answer you are looking for with a quick analysis on a small scale. It will save you time, energy, and resources.

15. USE .FORMAT() TO PRE-PROCESS TEXT DATA

Python string formatting method .format() is a very useful tool to pre-process text data. It can handle complicated string formatting quite efficiently. It has a built-in string class, which provides functionality for complex substitutions of variables and formatting of values (integer, floating-point, string, characters, etc.). The general syntax for using it is {}.format(value). Here's an example using the format method for single and multiple formatters.

```
#using format method in simple string with single parameter
Print("{}, Nurturing the world's top-league of data science professionals.".format("DASCA"))
```

The above program will give the following output:

DASCA, Nurturing the world's top-league of data science professionals.

```
#using format method with multiple parameters
my_string = "{} certifications for {} are available across {} countries."
Print (my_string.format("DASCA", "Data Scientists", 183))
```

The above program will give the following output:

DASCA certifications for Data Scientists are available across 183 countries.

16. TRANSFORM DATA INTO A NORMAL DISTRIBUTION

For data to follow a normal distribution pattern is of utmost importance for data scientists, especially if you are trying to do linear and logistic regressions. It is so because normally distributed data indicates the correctness of the researcher's assumptions, and makes the model more reliable. Additionally, most statistical calculations are done with the assumption that the data is normally distributed.

While there are numerous ways to test if your data is normally distributed using statistical software, here are a few transformations you can try on your numeric features to bring your data closer to normal or gaussian distribution using Python.

Transforming a distribution entails applying functions to the values such that it fixes the skewed data so that the output comes closer to what you expect. One of the most common goals of transforming the data is to bring it closer to normal distribution. Check out more on transforming data into normal distribution here.

- **Logarithm np.log(x) and Exponential np.exp(x)**
- **Inverse 1/x, square root np.sqrt(x), and cube root x**(1.0/3.0), and the likes**
- **Polynomial transformation like x**2, x**3, and so on**

17. PLOT DECISION TREES

Decision trees provide an intriguing way to extract value from data points. A decision tree is a sequence of yes or no questions, which eventually end in making a decision. It is a non-parametric machine learning modeling technique often used for classification and regression problems.

Say, a bank needs to decide if it should give someone a loan or not. To arrive at a decision, they will walk through a series of questions such as “What’s the income of the person?” If the answer is between $100 and $70,000, the bank could choose to continue to the next question, say “How long the person has been working in the current job?” and so on. Decision trees create a sequential and hierarchical decision process wherein decisional outcome varies with different data inputs.

Using sklearn, you can create a decision tree with just one line of code as shown below. The hyperparameters can be changed as per your needs.

```
#Sklearn function in Python to plot decision tree
plot_tree()
```

18. SCIKIT-LEARN IS MOST POPULAR FOR PREDICTIVE MODELING

Scikit-Learn (also known as sklearn) is a strong Python library with immense applicability for data scientists. It is used for machine learning and is built on Scipy, NumPy, and matplotlib libraries. This gives it functionalities in optimization, clustering, supervised learning, classification, regression, and many other techniques. It is perhaps the most useful library in Python. You must remember that the Scikit-learn library requires all the inputs in numeric, so you may have to convert your variables into numerical. Here are a few features Scikit-learn is popular for:

From linear regression, SVM, decision trees to Bayesian methods, the sklearn toolbox supports all types of supervised machine learning algorithms. Its use in supervised learning is one of the biggest reasons for its wide-spread usage.

Scikit-learn supports a gamut of unsupervised machine learning algorithms ranging from clustering, Gaussian mixture model, factor analysis, PCA to neural networks.

Sklearn offers many features and methods to select the metrics and check the accuracy of supervised models. It can be used for cross validation for evaluating estimator performance or plotting scores to evaluate models.

Scikit defines an API for creating visualizations. Some of the best features of this API are quick plotting and visualization adjustments.

The inspection module of sklearn helps understanding predictions from a model, and figure out what affects them. It can be used to assess your assumptions and biases about a model and design a better model.

Scikit-learn libraries can help you clean and preprocess your data, reduce it, expand it, and even extract features from texts and images.

By training a Scikit-learn model, you can make the model persist in the future without having to retrain it. You can do it with Python specific serialization and Interoperable formats.

19. IMPORTING PANDAS IN THE PROGRAM

Pandas has become a very popular library for data analysis. Not to mention, its huge role in ramping up the popularity of Python in data science.

It is used for data manipulation and structured data operations. Pandas has two data structures: Data Frames and Series. Series is a one-dimensional indexed and labeled array, which is used to access individual elements in your series with labels. Data Frames are columns and rows similar to Excel workbook.

Seasoned data scientists say that pandas data frames work like in-memory and in-Python data stores. So, if you have data that's big enough for your computations, but small enough for your computer, then pandas might just be right for you.

Importing Pandas in the program

```
# importing pandas in the program
import pandas as pd
```

```
# Defining a series object
srs = pd.Series([1,2,3,4,5])
```

```
# printing series values
print("The Series values are:")
print(srs.values)
```

```
# printing series indexes
print("\nThe Index values are:")
print(srs.index.values)
```

20. CREATE HEATMAPS ON PANDAS DATA FRAME

Heatmaps are simple. It represents numbers with colors. Say, higher numbers may be represented by warm colors in shades of red and lower numbers may be represented by cool colors in the shades of blue. Heatmaps have a better impact than conventional visualization techniques as they make spotting patterns easier. They will help you better understand your range of values in a glimpse.

Here's a simple code for creating heatmaps on pandas data frame using Seaborn

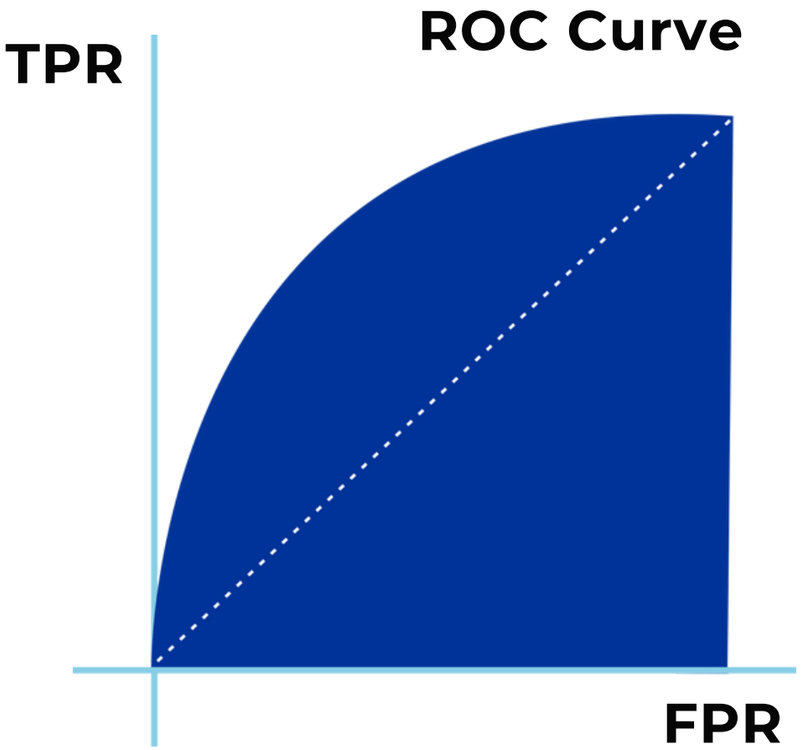
```
#import libraries
import pandas as pd
import mnatplotlib.pyplot as plt
import seaborn as sb %matplotlib inline
#define figure size for output plot
fig, ax = plt.subplots(figsize = (12, 7))
#define index and columns for data frame
idx = ['1', '2', '3', '4']
cols = list('ABCD')
#enter values in the index and columns as a panda data frame
df = pd.DataFrame([20, 30, 40, 50], [60, 70, 80, 90], [25, 35, 45, 15],
[5, 15, 10, 5]], columns = cols,index = idx)
#display dataframe as heatmap
sb.heatmap(df, cmap ='RdYlGn', linewidths = 0.50, annot = True)
```

21. PLOT ROC CURVES TO ENSURE MODEL INTEGRITY

ROC curve analysis is critical in both data science and statistics. It shows the performance of a model by assessing total sensitivity against the fall-out rate. Plotting ROC curve can reveal the model's viability. You may be surprised to know that it had immense applicability in WWII wherein it was used to detect aircraft of the enemy. Since then it has found uses in biology and other sciences.

Knowing the odds of corrections of predictions is among the main goals of data scientists. Since predictions are just guessing, a ROC curve analysis can give weight to your predictions. You can see how accurate are your model's predictions. In modeling algorithms, the ROC curve is typically the first test performed. It can detect problems early and let you know if you are on the right path.

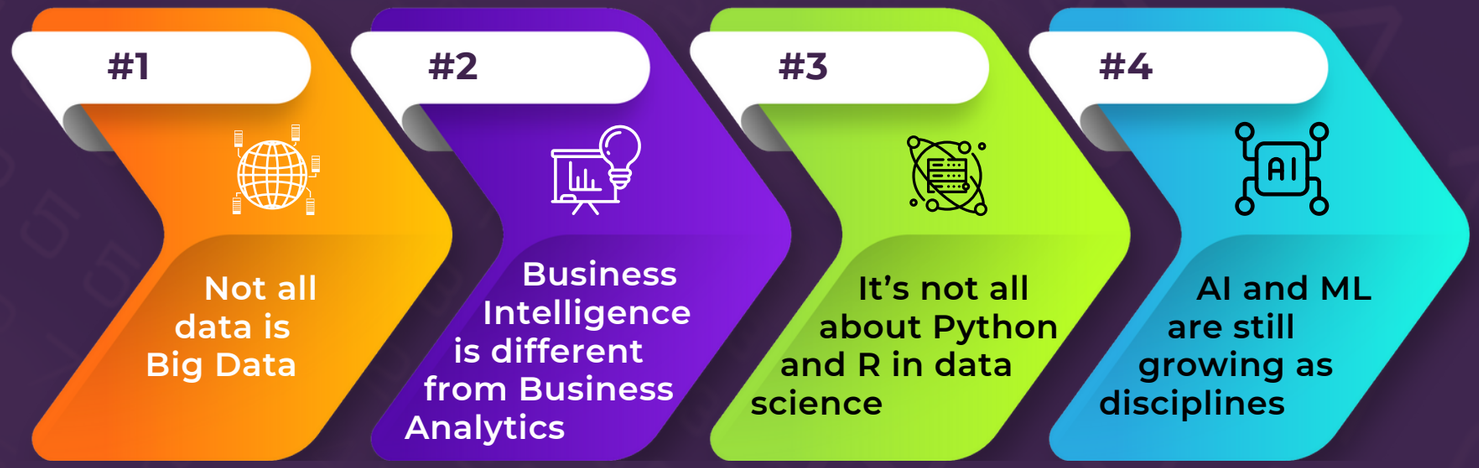
It plots the false alarm rate against the hit rate. ROC curve is a plot of the false positive rate (FPR) against the true positive rate (TPR) for a host of values between 0 and 1. FPR is calculated by dividing the number of false positives vy the total number of negatives (false positives and true negatives). Similarly, TPR is calculated by dividing the number of true positives by the total number of positives. These values are plotted on the curve with FPR on x-axis and TPR on y-axis. ***The further the curve from the line, the more predictive is your model.***



DATA SCIENCE COUNCIL OF AMERICA

DEBUNKING COMMON MISCONCEPTIONS

With the overflowing bandwagon of data science, many do not yet fully comprehend the frequently used terms and buzzwords in the field. It has led to the proliferation of several myths and misconceptions among data science professionals. Here's to refute and clarify some of the most common misconceptions.



DATA SCIENCE COUNCIL OF AMERICA

1 Not all data is Big Data

Calling 3000 lines of data as Big Data is not correct. You must consider several factors before you dub a data set as Big Data. This is important to understand as the kind of data you think you are dealing with also has a bearing on the tools you will choose for analysis. Volume is just one of the parameters of defining Big Data. It is also characterized by the types of data that are involved (variety) and perhaps the variability of the sources. Additionally, the velocity with which the information is being received and being processed. Therefore, never dub a data set as Big Data just by one factor alone.

2 Business Intelligence is different from Business Analytics

A few understand this, but there is a significant difference between business intelligence (BI) and business analytics (BA). They are not interchangeable terms. Furthermore, not all data analysis can be considered business intelligence. BI deals with the explanation of past events and future predictions using data-driven approaches and reaching data-based conclusions.

DATA SCIENCE COUNCIL OF AMERICA

3 It’s not all about Python and R in data science

Data Science is often associated with strict coding practices and super quantitative skills. It’s widely believed that data science is all about mastering R and Python. Although, it is not always so.

Software like Excel, STADA, SPSS, and Tableau are quite frequently used by them. Sometimes, even more than R and Python. Storytelling is among the many other crucial skills of data scientists. All data scientists must have the stomach to process and express complex mathematical and programming concepts in a story-like presentation to the stakeholders. So, if you understand data, know statistical software, and are good at storytelling, then to begin with you could write code using libraries and you would still be good to go. Data Science, in essence, is a combination of many skills – from mathematics, programming, and business. Overemphasizing just one would be a mistake.

4 AI and ML are still growing as disciplines

The talks around Artificial Intelligence and Machine Learning gather much attention. However, you must understand that they are not so old disciplines. The lightning progress in these domains is a phenomenon of very recent developments, and many developments are yet to happen. Scientists, academics, and researchers still debate at length about basic functionalities. To give an example, it is widely known that deep learning can make a model perform exceptionally well. Albeit, it is still not clear how a machine can obtain those levels of results given existing capabilities. So, as you use ML and DL algorithms and models, know that there is much that we don’t know.

DATA SCIENCE COUNCIL OF AMERICA

GROWING YOUR SKILLS

The secret to a great data science career is incessant learning, rich qualifications, and advanced skillsets. If you are looking to take your career up by several orbits, here are some ways to gain the relevant skills and knowledge to successfully grow in and demonstrate your fitness for data science

DATA SCIENCE COUNCIL OF AMERICA

1 Bootcamps and online courses

The pre-requisite skills for data scientists fall under three categories – programming and hacking, mathematics and statistics, and business acumen. You can develop your competencies by leveraging online resources such as video tutorials and online courses and attending bootcamps. Bootcamps offer a fast-tracked intensive method of training. You could also learn by reading books authored by seasoned and successful data science professionals.

---

2 Professional certifications

As you start finding jobs, you'll realize that demonstrating your skillsets for data science during the interview is tougher than it seems. Given the scores of competitors eyeing for the same job, you will need to stand out. Earning a professional certification in data science can demonstrate your promise and potential for the job, and make you more noticeable to the hiring teams.

Professional certifications are designed to assess aspirants on strict professional standards. Acing them speak volumes about your fit for the job role. DASCA offers three rungs of prestigious international credentials in data science for Data Engineers, Data Analysts, and Data Scientists. Depending on where you are in your professional journey, they can prove to be great career-shaper qualifications for you. You can contact us to find which qualification will suit you the most. Explore more here.

One thing I would strongly emphasize is that you need to demonstrate you can do this job. Some ways I have seen people demonstrate their skills and capabilities is through open source contributions, speaking at local meetups on projects they've done, and developing a portfolio of projects. For me, I took a combination of all.

JULIA SILGE

Data scientist and software engineer, RStudio

DATA SCIENCE COUNCIL OF AMERICA

3 On-the-job training

Experience plays a big role in getting a job. The best way to get it is by training yourself in data science while you work. Since work is where we spend most of our time, you can find the projects at your company that may benefit from a data-driven approach. As you take them up, it will help you gain on-the-job expertise in data science. Check if your leadership is encouraging toward it. You could also assess if your company can sponsor for you formal training to develop your skills. This route, of course, will mean more workload than usual, but if you are motivated enough, both you and your company can benefit from this arrangement.

4 Building portfolio

Networking with fellow data scientists and building your own data science projects can make a world of difference in your professional standing. They both demonstrate your passion for learning and can give you an edge over others. You can start by making a GitHub profile. Every strong portfolio has two parts: GitHub repo, which hosts the code for the project, and a blog, which explains what you did and why, and gives you an opportunity to show off your communication skills.

Additionally, if you make your code public, you can interact with other data science enthusiasts. You can collaborate with them on a few pet projects and also seek
guidance from your peers about how to grow in the data science field. Kaggle is another popular data science community platform you can explore.

ABOUT DASCA

The Data Science Council of America (DASCA) researches, designs, and builds platform-independent Data Science knowledge frameworks, standards, and credentials, and certifies individuals entering or working across the spectrum of emerging Data Science professions. The prime goal of DASCA is to develop high-quality professionals who can squarely address the challenging expectations of Data Science stakeholders internationally.

DASCA offers the world's most powerful set of credentials along three critical profession-tracks in data science – Big Data Analytics, Big Data Engineering, and Data Scientists.

DASCA credentials validate the promise and potential of professionals to hit the ground running in the most demanding assignments and roles. DASCA-certified individuals bring to the table unmatched understanding and capabilities to anticipate and appreciate the need for deploying the latest Data Science techniques, tools, and concepts to manage and harness Big Data across verticals, environments, and markets.

dasca™

DATA SCIENCE COUNCIL OF AMERICA

For more information, please visit www.dasca.org