CODEUP'S GUIDE TO DATA SCIENCE

Since Harvard Business Review named Data Scientist the "Sexiest Job of the 21st Century" and Glass Door ranked it the #1 Best Job in America, the buzz around Data Science has exploded globally. In fact, Google searches for the term have quadrupled in the last 5 years. Along with the instant fame, a lot of questions have come up. What exactly is Data Science? What are the most common tools and technologies? How do you become a practitioner?

Codeup is here to answer all your Data Science questions !

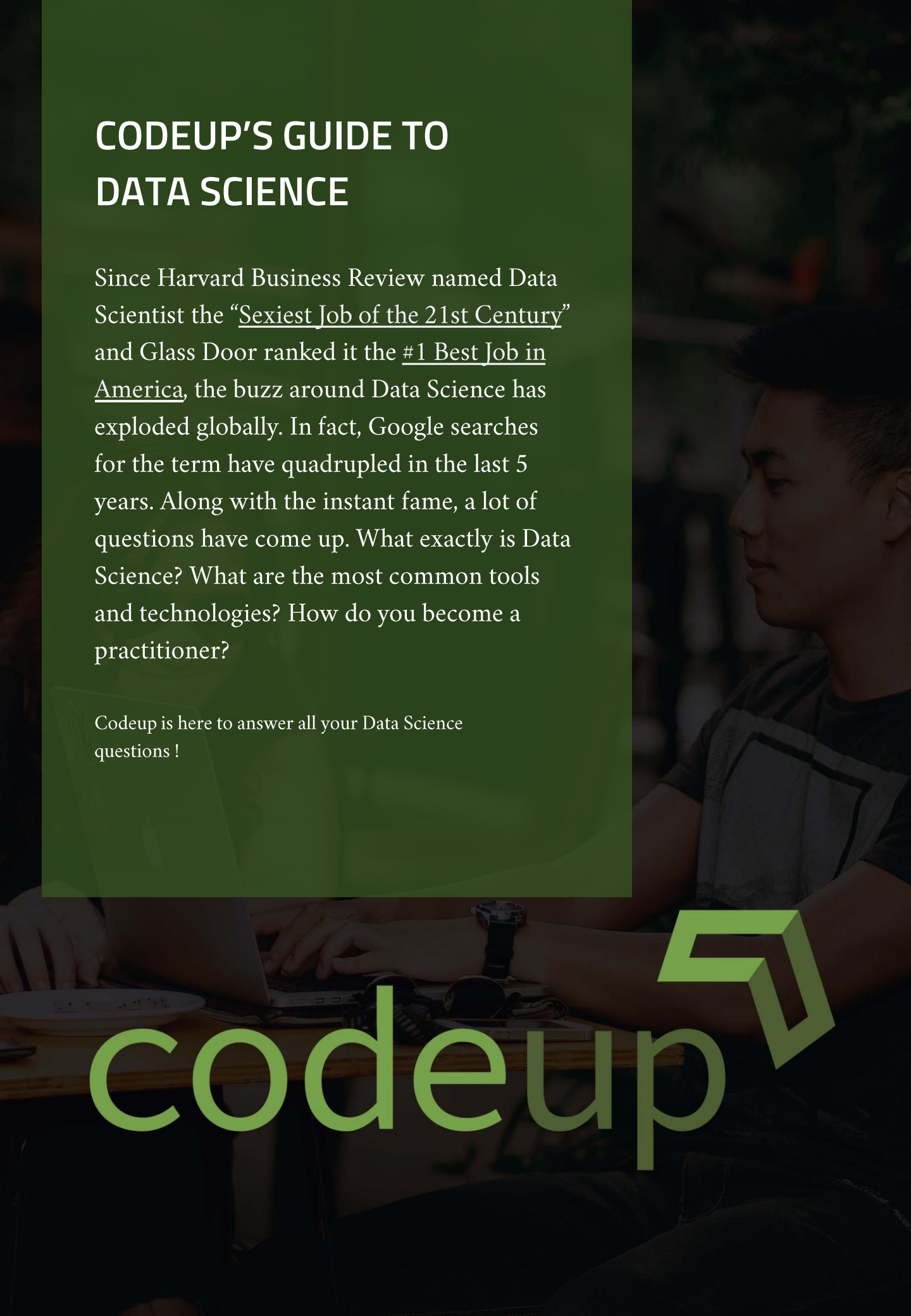


TABLE OF CONTENTS

WHAT IS DATA SCIENCE ?

First of all, data science is a method of providing actionable intelligence from data using math, statistics, programming, and business expertise. Like any scientific method, it involves gathering data, identifying a problem, forming a hypothesis, and running tests. More specifically, data scientists follow a pipeline process of data acquisition, wrangling, exploratory analysis, model development, product delivery, and storytelling. Practitioners typically spend 70-80% of their time in the wrangling/exploration, 20% on machine learning models, and the rest in maintenance. Most importantly, this whole process should result in a valuable action or insight for the end-user, i.e. a business or customer!

At its core, data is just information – names, dates, times, $$, etc. Data scientists work with large collections of this information to draw conclusions. For example, they might use financial data to predict seasonality of revenue generation or use the events of applications (like logins, clicks, or downloads) to detect security threats or fraud. Data science typically deals with 'Big Data,' which is too large and complex to manage on a local computer. People interact with and create data like this every day: using smartphones, buying houses, rating movies, and more. You can thank data scientists (and the teams supporting them) for guiding you to your favorite Netflix series and helping optimize your workouts.

There's always more than one way to eat an oreo, and Data Scientists use dozens of different tools. On top of that, they work with concepts from math, stats, and programming to write functions, create charts and graphs, and model patterns. At Codeup, we modeled our tools after industry standards, which include Excel, Python (programming language), SQL (databases), Spark (distributed data), Jupyter Notebooks (virtual notebooks for doing data science), and Tableau (visualizations).

WHAT ISN'T DATA SCIENCE? ANALYTICS AND OTHER MYTHS

Now that we know what Data Science is, what isn't it? More specifically, one of the most common questions that we see is: what is the difference between data science and data analytics?

First, let's redefine some of our terms! Data science is a method of turning raw data into action, leading to a desired outcome. Big Data refers to data sets that are large and complex, usually exceeding the capacity of computers and normal processing power to deal with. Machine Learning is the process of 'learning' underlying patterns of data in order to automate the extraction of intelligence from that data.

Now, let's look at the data pipeline that data scientists work through to reach the actionable insights and outcomes we mentioned:

- We start by collecting data, which may come fro social media channels, network logs, financials, employee records, or more.
- We then process that data into usable information stored in databases or streamed.
- Next, we look back on the history of that data to summarize, describe, and explain,turning the data into meaningful knowledge. Here we'reprimarily using mathematics, statistics, and visualization methods.
- Now we convert that knowledge into intelligence,seeking to predict future events so that we can make decisions in the present. This is where practitioners will introduce mathematical/statistical modeling through machine learning to their data.
- Finally, we enable action by building automations, unning tests, building visualizations, monitoring new data, etc.

Data professionals work at different stages of the spectrum to move data through the pipeline. On the left, Big Data Engineers specialize in collecting, storing, and processing data, getting it from Data to Information. In the middle, analysts work to understand and convert that information to knowledge. Lastly, a Machine Learning Engineer utilizes machine learning algorithms to turn intelligence into action by building automations, visualizations, recommendations, and predictions.

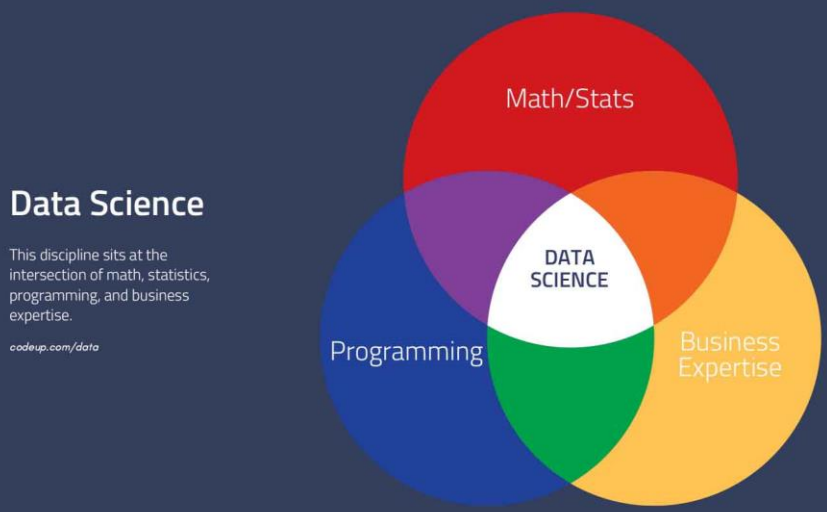
Data Scientists span multiple stages of this pipeline, from information to action. They will spend about 70% of their time wrangling data in the information stage. They will conduct statistical analysis to derive knowledge. Lastly, they predict future events and build automations using machine learning. For those technical folk out there, data science is to data engineering or machine learning engineering as full-stack development is to front-end or back-end development. For the non-technical folk, data science is the umbrella term that houses data analytics, machine learning, and other data professions.So what's the biggest difference between a data analyst and a data scientist? Data scientists utilize computer programming and machine learning in addition to mathematics and statistics. Moreover, they often deal with bigger, messier, and live data.

Now, let's tackle some common myths and misconceptions in the field of Data Science. Data Science, Big Data, Machine Learning, NLP, Neural Networks...these buzzwords have rapidly spread into mainstream use over the last few years. Unfortunately, definitions are varied and sources of truth are limited. Data Scientists are in fact **not** magical unicorn wizards who can snap their fingers and turn a business around! Today, we'll take a cue from our favorite Mythbusters to tackle some common myths and misconceptions in the field of Data Science.

Myth #1:
Data Science =Statistics

At first glance, this one doesn't sound unreasonable. Statistics is defined as, "A branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data." That sounds a lot like our definition of Data Science: a method of drawing actionable intelligence from data.

In truth, statistics is actually one small piece of Data Science. As our Senior Data Scientist puts it, "Statistics forces us to make assumptions about the nature of the relationship between variables, the distribution of the data, etc." In the traditional Data Science venn diagram, you'll see that math/stats make up ⅓ of a working professional. These are tools and skills to leverage, but data science itself is about drawing intelligence from data.



Myth #2:
Data Science = Data Science

This one's tricky, because it's impossible to either confirm or bust! The 'myth' is that one person or company using the term Data Science is not necessarily the same as ***another*** person or company using the same term. Depending on organizational capacity, individual experience, educational background, and many other variables, we might be using the same name for different animals.

Tl;dr: don't assume a common understanding across hiring managers, recruiters, and practitioners. Look instead for specifics of tools, techniques, methodologies, and outputs. That being said, this one falls in the "plausible" category, because it may actually be true in some circumstances, while false in others.

Myth #3: Data Science curricula are well-defined and consistent.

We recommend checking this one out for yourself! A quick google search for bootcamps, master's degree programs, and online courses will reveal that different organizations teach different things. There is no commonly accepted framework for teaching data science! Some focus more on the engineering, others focus more on machine learning, some think deep learning is foundational, and some prefer to use R.

Our curriculum was built through employer interviews, practitioner interviews, market research, and company partnerships. But we're based in San Antonio! A bootcamp in New York might follow the same process and end up with a different syllabus. Keep in mind, whatever your learning path, that there will be gaps in your learning. The most important thing is to recognize those gaps.

MYTH #4: IF I WANT TO BE A DATA SCIENTIST, I JUST NEED TO LEARN PYTHON OR R.

This one is common and dangerous! Just like statistics, programming languages like Python and R are *tools*. They're just pieces of a larger puzzle! Knowing Python without understanding the data science pipeline is like knowing how to build a floor without having a floor plan. Of course, these are valuable technical skills that give you a leg up, but they're second in importance to asking the right questions, knowing what tools to use when, and communicating your findings.

WHERE DO DATA SCIENTISTS COME FROM ?

Now that you know what it is and what it's not, what can you do? If you're interested in becoming a data scientist, you might be wondering how other people got into the field. Given how new the profession is, most of today's practitioners probably didn't study data science formally as undergraduate or graduate students. So today we're asking: *where do data scientists come from?*

Let's start broadly by defining the possible pathways into this career. If you're a Data Scientist, you probably followed one or more of these paths:

- Learning on the job: You 'did it live' and hacked your way into a data science skillset.
- Universities: You studied Data Science, Analytics, Statistics, Programming, or Business formally in a university setting.
- MOOCs (Massive Open Online Courses): You learned through an online resource like Udemy or Codecademy.
- On-site or corporate training: You were trained by a learning & development department, internal academy, or contracted provider.
- Immersive programs/bootcamps: You went to coding bootcamp and learned Data Science through an immersive, hands-on career accelerator (like Codeup, perhaps?)

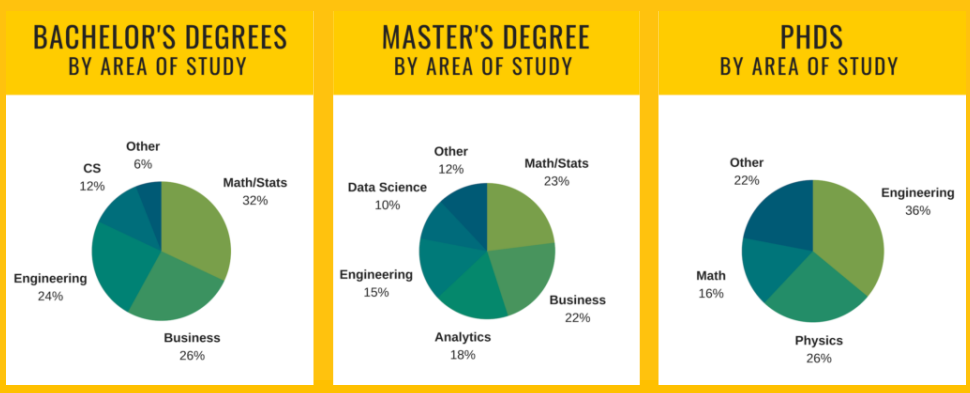
Each of these pathways has unique advantages *and* disadvantages across variables like cost, formal credentials, length, and pace. A free online program is free and accessible, but takes a lot of dedication to follow through and is harder to change careers with. A bootcamp specializes in quick and efficient job outcomes, but is a big investment. A university offers a formal degree and dives deeper, but is more expensive and takes longer.

Each of these pathways also leaves gaps against the complete picture of a data scientist. From your training you might be missing components like: working with real data sets, understanding industry and company demands, using up-to-date technologies, or even just knowing what you don't know! What's important to understand here is that different pathways yield strengths and gaps. Your job is to find, acknowledge, and improve your gap areas!

Now that we have a framework for understanding potential pathways, let's look at some data. In preparing to launch Codeup's immersive Data Science program, we researched over 250 data scientist profiles on Linkedin and analyzed their educational and career histories. Here's what we found!

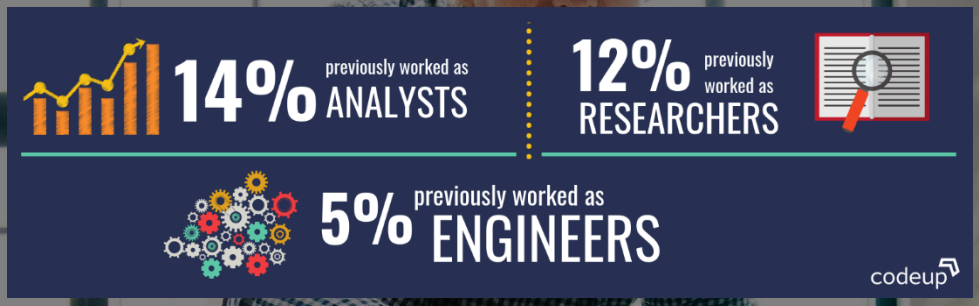
EDUCATION

95% have a Bachelor's Degree, 70% have a Master's Degree, and 27% have a PhD. Of those degrees, the most represented areas of study are: math, stats, business, engineering, and CS.



CAREER HISTORY

Data Scientist come from an incredibly diverse range of professional backgrounds: psychology research, software development, business analyst, mechanical engineering, and more! We saw a few prominent patterns in our data:



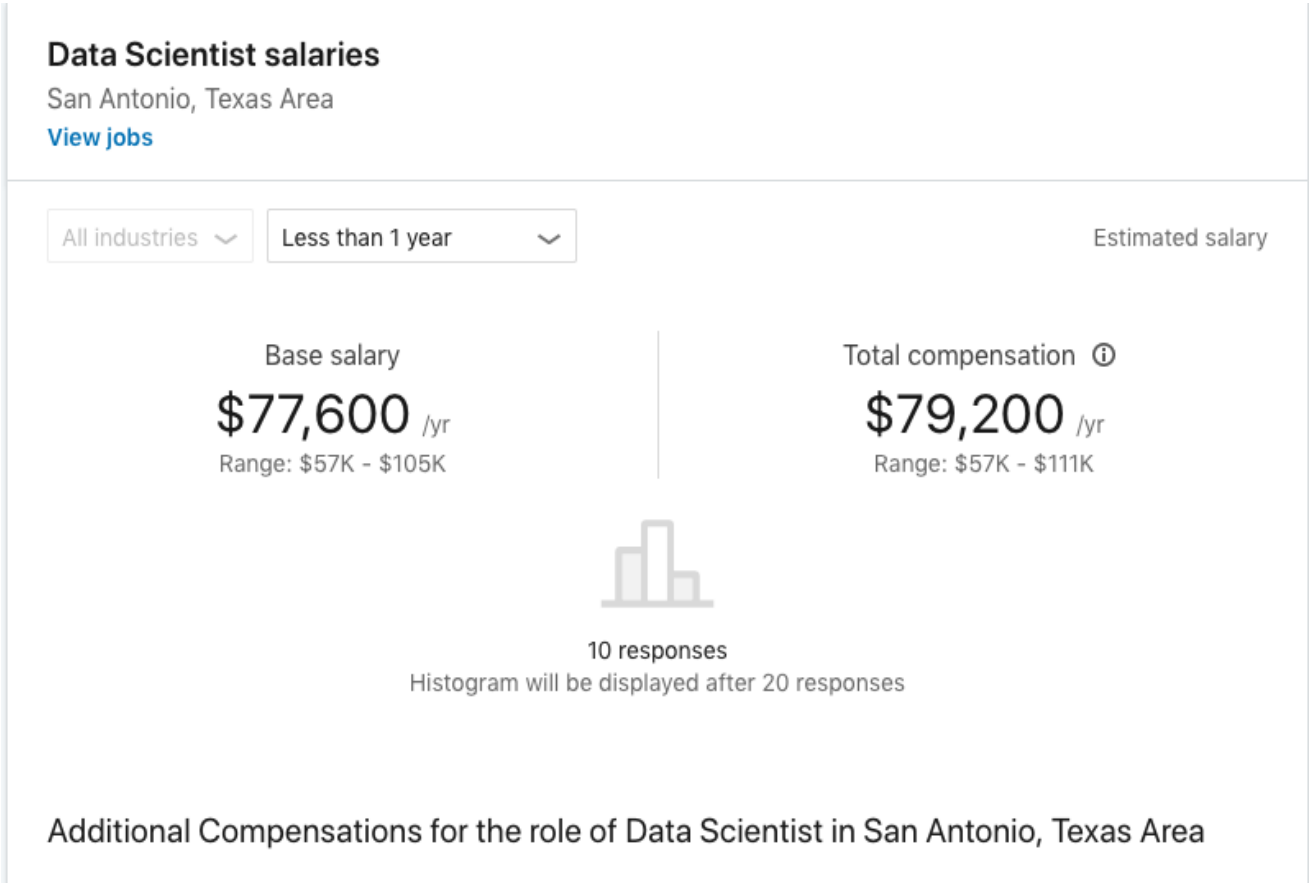
ARCHETYPES

A third component of our research was to interview practicing Data Scientists. We asked questions like: What was your path to the field? What did you study? Is there a need for programs like Codeup? What are the most important skills to learn? After conducting these interviews, we had three valuable lenses to understand the paths into Data Science: educational histories, career histories, and first-hand qualitative research. From these three, we compiled 4 archetypal Data Scientist personas!

Why Should I Care?

Do you enjoy your curated Netflix movie queue? Do 50% of your Amazon purchases come from "Customers also bought…" recommendations? Does HEB hit you with the perfect coupon combo? Does it make you feel safe that your bank calls you when they suspect fraudulent activity? If the answer to any of those questions is yes, you have Data Science to thank!

The scene in San Antonio is relatively new, so there's a lot of opportunity. Not only is this an in demand skillset nationally, but it's a well paid one even right here in San Antonio. Reports show the average salary for a Data Scientist with 0-1 years of experience to be $77,800.



HOW DOES CODEUP FIT IN?

Codeup launched its first training course teaching full-stack web development in February 2014. Since then, we've reverse engineered a full-stack curriculum directly from the jobs our hiring managers recruit from us for. Our model is built around immersive in-person learning with a job-focused curriculum and outcome focused-process. So when we began hearing from our partners that they were struggling to hire Data Scientists, we hit the streets to find out more.

To begin with, we researched the field from 3 angles: job opportunities, first-hand interviews with industry professionals, and curriculum research. We wanted to find out if this would be a worthwhile program offering. Glassdoor, indeed, and our hiring partners all told us the opportunities were plenty. Check! Next, we interviewed over 30 practitioners and hiring managers to find out what they were looking for. They were looking to hire, and we knew what they wanted. Check! Lastly, we researched other programs - bootcamps, Master's, Undergrad, and more. We discovered that there was some overlap in the content being taught, but most programs were piecemeal, not delivering a cohesive and linear curriculum. Additionally, many required tons of programming experience to even get in.

With the needed verified, some data in hand, and a gap identified, we set out to build the best data science curriculum out there. We begin with 40 hours of pre-work to get students situated in the industry, programming languages, and math/stats skills they'll need. In class, we teach the full data science pipeline from acquisition through storytelling. We vary our data sources and tools, and leverage real world and messay data to replicate on the job work. We start with the basics: excel, data analytics, math, and stats. Then we build a foundation in the two most important technical skills: Python and SQL. Next, we practice a range of applied methodologies, including classification, natural language processing, anomaly detection, and clustering. We next leverage Spark to learn how to work with distributed data in the cloud. Finally, we wrap it all up with some data storytelling and visualization using Tableau.

WHAT NEXT?

So now you know what Data Science is and what it's not. You've busted some industry myths. You understand the various pathways into the field and backgrounds of practitioners. You know some of the real world applications of the skills and understanding of the urgency and need. And lastly, you know how Codeup approaches training the next generation of Data Scientists. So what's next? You are! If you're interested in a career in Data Science, reach out to us at info@codeup.com - we'd love to help you create your tomorrow.


codeup