Data science has become an ubiquitous field in the 21st century, finding a use in practically all other existing fields of study. However, the fact remains that the complex jargon associated with data science can only be understood by data scientists. I believe that it is fundamental for data scientists, both new and veteran, as well as laymen to understand the jargon in order to facilitate communication between data scientists and the businesses they assist. I will be primarily defining the terms that a layman would not be familiar with (i.e., not simple terms such as *algorithm* or *database*). I will also define some Python terms unfamiliar to a layman, as Python work composes a good chunk of data science as well.

**Data Science Terms:**

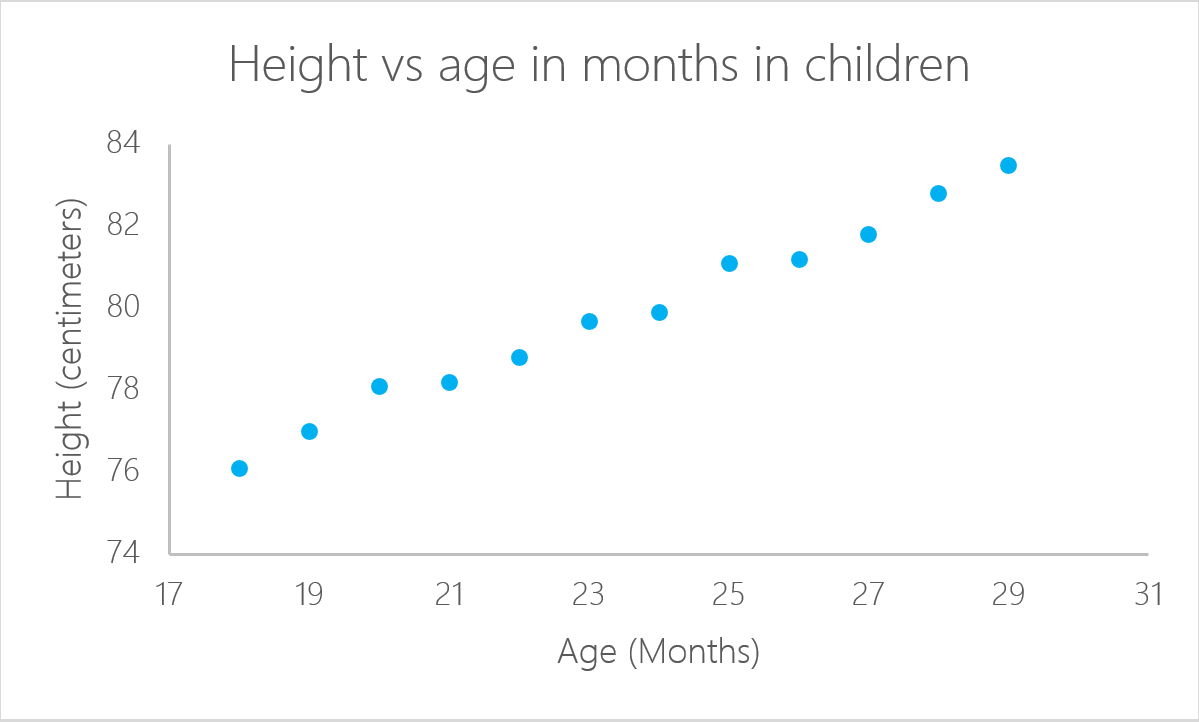
Data Warehouse: A data warehouse is a type of database, a library used for storing data. Both current and historical data collected by a company or organization are stored here to be used for making analytical reports; this helps make important business decisions.

EDA: An acronym that stands for Exploratory Data Analysis, EDA involves summarizing the characteristics of a dataset to determine any logical next steps, either via graphs/plots or statistics. For example, if our goal was predicting car CO2 emissions, we might want to start with a graph of CO2 emissions versus many other variables to get an idea as to what variables will influence emissions the most. Below is an example of such a graph.



ETL: Standing for Extract, Transform, Load, this describes the three steps used to blend data from multiple sources. For example, let’s say we had data from different sources all sharing one column, but all the sources use different units and scales. To consolidate these data, we would first have to convert them to the same units and use a process called normalization to get them into the same scale. Only then can we merge the data sets.

Regression: Regression determines the relationship between two variables on a graph by fitting a line that best fits the point. The equation of this line doesn’t have to be linear, it could be quadratic The equation of the line can be used to predict unknown Y values from known X values. For example, let’s say we have data points of height versus age (shown below). We can generate a line that best fits the data; the equation of this line should predict a child’s height (Y value) from their age (X value).



**Python Terms:**

Function: Most people would think of a function as an algebraic equation. Functions in Python can be equations, but they don’t have to be. Rather, they are chunks of code that perform a function, hence their name. They’re useful for saving repetitive lines of code. For example, one could program a function to do math, to read a file, or to find a word’s frequency in some text. The latter is exactly what Ctrl + F does on a computer.

Dictionary: A dictionary in Python isn’t a book of definitions. It’s a set of key-value pairs, where one key is assigned to one value. These keys and values could by alphabetic, numerical, or both. They’re useful for storing data online, such as customer information and login credentials. Below is an example of a dictionary. Each key-value pair is separated by a colon.

Dict = {‘Name’ : ‘Bilal’, ‘Gender’ : ‘M’, ‘Height’ : 173}

Loop: A loop in all programming languages is an instruction to keep running a chunk code until a certain condition is met. For example, if we wanted to run a piece of code 5 times, we would specify “for each number in 1 to 5, execute a certain command” as computers do *exactly* what their instructions say. Of course, this would be in the respective programming language.

Float: No, this is not a life-saving aquatic device! It is short for a floating-point number, or a decimal number. Python treats them as their own data type separate from whole numbers, so 5 and 5.0 are classified differently despite producing the same mathematical output.

References:

Bakthavachalam, V. (2019). *25 Terms Every Data Scientist Should Know.* Coursera. Retrieved on December 1, 2020 from https://blog.coursera.org/ds-academy-25-data-science-terms/.

Dedic, N. & Stanier, C. (2016). *An evaluation of the challenges of Multilingualism in Data Warehouse Development.* Staffordshire University. Retrieved on December 1, 2020 from http://eprints.staffs.ac.uk/2770/.

Kudaimi, B. (2020). *DSC 530 Final Project: Predicting Vehicle CO2 Emissions from Engine Specifications.* Cyberactive Blackboard.

Sims, Z. & Bubinski, R. (2020). *Python Glossary*. CodeAcademy. Retrieved on December 1, 2020 from https://www.codecademy.com/articles/glossary-python.