An In-depth Analysis of Careers in Data Science: A K-Means Clustering Approach

Emergent Research Forum (ERF)

Lakshmi Iyer
Appalachian State University
iyerls@appstate.edu

Mitchell Church
Coastal Carolina University
iyerls@appstate.edu

Daniel Emery
Appalachian State University
emeryde@appstate.edu

Abstract

The combination of a significant talent gap in the data science industry (Lyon and Brenner 2015) with increasing demand for data professionals in this relatively new field has created a market in which there is more work than there are qualified people to fill positions. Past research has identified two major causes of this problem. First, conflicting job descriptions have created significant ambiguity across seemingly similar job categories (Carter and Scholler 2016). While data scientists, data engineers and data analysts seem similar in both name and job description, the actual work typically asked of each job category can differ significantly, and these differences are becoming starker and more pronounced (Kim et al. 2016). Secondly and perhaps more importantly, companies hiring practices have failed to keep pace with changing attitudes and expectations across job categories (Waller and Stanley 2013). As a result, many job descriptions contain elements of multiple job categories (3 to 8 as seen in industry reports from PWC, Forbes.com, Udacity), conflicting task expectations, and uneven expected salary reporting. This all leads to confusion among applicants and dissatisfying outcomes for employers.

Another challenge relates to the education of individuals for data roles. As the need for data related professions has increased, educational institutions have developed curriculum and programs designed to give students the skills necessary to succeed in the changing job market (Iyer and Schiller, 2014; Schiller et al. 2015). Business analytics programs throughout the company exhibit demand that consistently outpaces supply (Turel and Kapoor 2016). However, many of these curriculum suffer from the same ambiguity present in job descriptions discussed above, leading to programs that present an amalgam of many different data jobs rather than one coherent vision or desired field (Jacobi et al. 2014). Such issues result in graduates who may have a difficult time convincing employers of the value of their education, or in worst case genuinely being unprepared for any job in a data field today.

Our study seeks to address these concerns and fill a current gap in research for providing some clarity into common trends and best practices related to the development of job descriptions. About 20,000 data science job postings were collected from Dice.com using a custom web-scraper built in Python 3.6. K-means++ clustering and “elbow” method revealed the five most important job categories as analyst, data architect, data engineer, data programmer, and data scientist. Results of the study show that the these job categories vary significantly in the way they advertise the required skills of each position. Data scientist listings were most likely to demand a consistent skill set around big data. At the same time, analyst and data architecture positions suffer from significant ambiguity in the way that their work is described. As part of future research we will analyze the data along with industry findings to further refine the categories and provide useful insights to inform students and job seekers in data science.

Keywords: Data Science Careers, K-Means Clustering, Data Analyst, Data Engineer, Web Scraping.