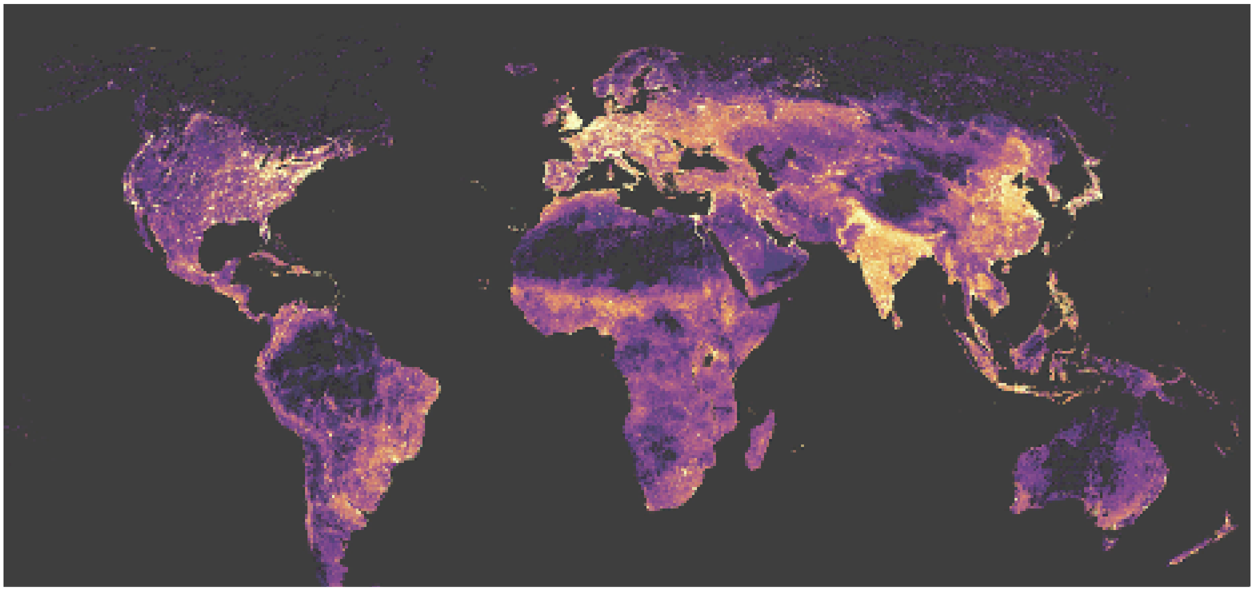

November, 2019

Adrian Lievano

Table of Contents:

Purpose: Provide insights to build and support data-teams.

© 2019 Adrian Lievano All Rights Reserved

What’s your data culture?

Motivation:

Data is everywhere -- in every industry, country, organization, and user of digital applications, data and the way we store, process, analyze, and share its insights with others can be used for great benefit. Leaders across companies and prospective job seekers interested in information are on fertile grounds: the cost of data storage is exponentially decreasing, the amount and velocity of data is increasing, and the algorithms that open the valve on this spigot of value are more accessible with modern programming frameworks. To capture this value, however, companies face considerable challenges such as hiring and retaining talent, using an organization’s structured and unstructured datasets, and much more. The best way to tackle these problems is to have a data strategy: a strategy for organizing, governing, analyzing, and deploying an organization’s information assets[11].

A data strategy has multiple parts: addressing compliance and security, creating new products and services, or developing organizational analytics capabilities to name a few. A crucial element in creating an effective data strategy, however, starts by creating your data culture; it influences the competitive advantage when your bring talent, tools, and decision making together. There are multiple surveys of c-suite executives from various Fortune 500 companies, each adding a unique understanding of the makings of a strong data culture [16]. In this report, however, we add to the conversation by providing insight into building technical teams and how understanding the nuances of your data defines your data culture. As a result, I aim to empower executives with the insights to build data-driven cultures.

Background:

Companies that prioritize data-driven decision-making create competitive advantages in their industries: Lyft, Didi Chuxing, Facebook, Google, Apple, among others, are examples of the most valuable businesses that leverage data and analytics to create new products, improve on existing products or services, or attract talent [X]. Despite the economic opportunities present in data across industries, progress towards creating data-driven cultures is stagnant: of 64 surveyed c-level technology executives at some of the largest corporations, 72% report that they do not have a data culture, 69% are not data-driven, 53% are not treating data as an asset, and 52% do not believe they are competing on their data assets and analytic capabilities. In attempts to address these issues, a staggering 93% of respondents identify people and process issues as the main obstacle [X]. For companies to benefit from their data, they need to become a data-driven organization and a great first step starts with understanding the people in a data team.

In the modern data science team, roles include machine learning engineers, data engineers, data scientists, product managers, analysts, and software engineers. The roles overlap, and vary in programming, mathematical, and communication proficiency, but each contribute crucial skills to use data to accomplish business goals.

© 2019 Adrian Lievano All Rights Reserved

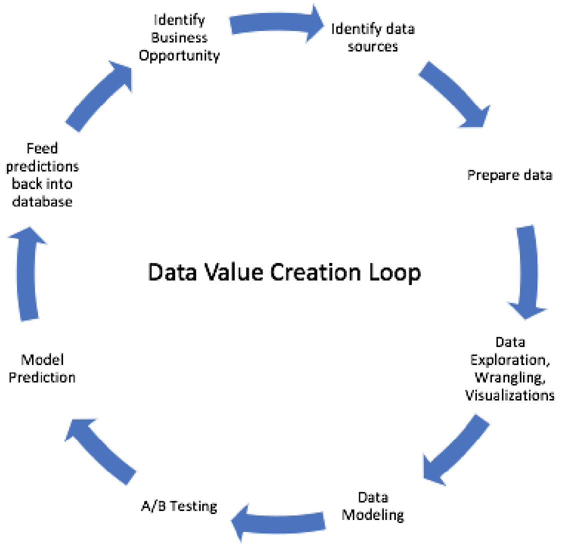
Identify Business Opportunity
Identify data sources
Prepare data
Data Exploration, Wrangling, Visualizations
Data Modeling
A/B Testing
Model Prediction
Feed predictions back into database
Data Value Creation Loop

Exhibit: A simplistic depiction of the data-value creation loop.

A business goal using data can be achieved, for example, by following a data value creation loop: a sequence of well-defined steps that involve generating revenue from data. Awareness of the data-value creation loop and its potential impact on a company's revenue does not fall on deaf ears: 92% of the c-level respondents reported an accelerating rate of investment into "artificial intelligence" and 55% of them report investments in Big Data and AI exceeding $50MM and growing. There is a misunderstanding, however, because increasing investment dollars into AI without the foundation in the data value creation loop in place can have serious consequences -- it's putting the cart before the horse.

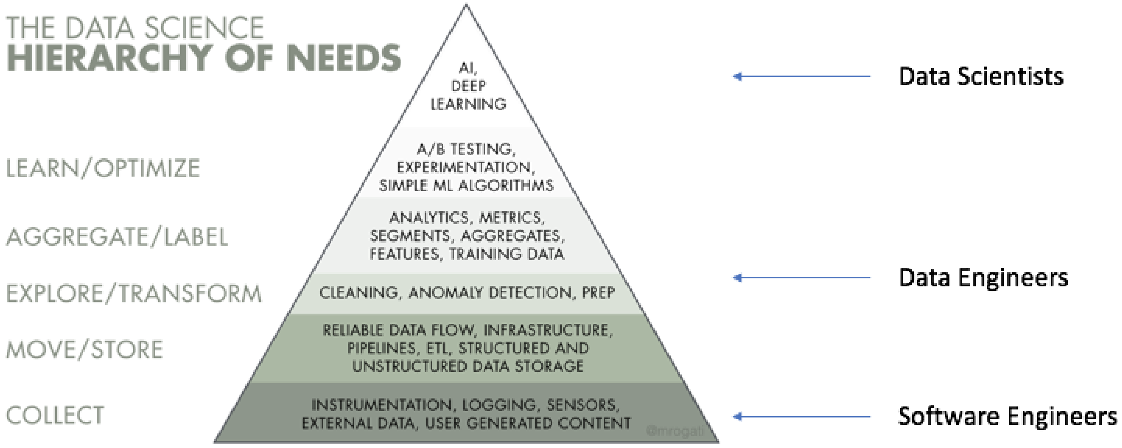
THE DATA SCIENCE
HIERARCHY OF NEEDS
LEARN/OPTIMIZE
AGGREGATE/LABEL
EXPLORE/TRANSFORM
MOVE/STORE
COLLECT
AI, DEEP LEARNING
A/B TESTING, EXPERIMENTATION, SIMPLE ML ALGORITHMS
ANALYTICS, METRICS, SEGMENTS, AGGREGATES, FEATURES, TRAINING DATA
CLEANING, ANOMALY DETECTION, PREP
RELIABLE DATA FLOW, INFRASTRUCTURE, PIPELINES, ETL, STRUCTURED AND UNSTRUCTURED DATA STORAGE
INSTRUMENTATION, LOGGING, SENSORS, EXTERNAL DATA, USER GENERATED CONTENT
@mrogati
Data Scientists
Data Engineers
Software Engineers

Exhibit: You need a solid foundation for your data before being effective with AI and machine learning. Credit: https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007

At the bottom of the pyramid, software engineers interface with sensors located in devices (mobile devices, industrial machinery, etc.) to collect data. User-interfaces provide a medium for consumers or enterprises to interact with software and provide user-driven information that provide insight into their usage patterns. Data engineers interface with these unstructured data in a variety of formats and program processing algorithms to extract, transform, and load the data into structured, accessible formats. It is at this point where more value can be captured — for example, analysts or data scientists can gather sample statistics, clean the data, or build visualizations to inform strategic initiatives. In the explore and transform part of the pyramid, dashboards can be presented to cross-functional teams and provide actionable insight based on company data. At the learn and optimize level, data scientists either develop their own machine learning models or work with machine learning engineers to design experiments. At the top -- the level of where most of the corporate investment dollars go towards -- artificial and deep learning technologies can be applied.

<add paragraph to transition into next key points>

Methodology:

The annual industry-wide Kaggle Data Science & Machine Learning survey contains 16,000, 23, 859, and 19,717 responses in 2017, 2018, and 2019, respectively[2, 3, 4]. A Kaggle data science notebook and jupyter notebook is used to analyze the survey fields. This report focuses on self-reported Software Engineer, Data Engineer, and Data Scientist respondents. I selected this audience because these are the key contributors in a data team: software engineers build the infrastructure that allow user actions to be logged, data engineers extract, transform, and load (ETL) these actions into structured tables, and data scientists use this data to analyze, predict, or communicate results to various stakeholders. It’s important to understand their different needs so that organizations that seek to build a data-driven culture can invest in key contributors to solve their major obstacles. All code, visualizations, and supporting resources can be found in the reference section of this notebook. I also cite external studies, referenced below.

Discussion:

Section 1.0 - Understanding Technical Contributors: Who are they and what do they do?

1.0 Purpose: Why do we care?

A shortage of the analytical and managerial talent to leverage data is an obstacle companies can begin to face in the short term. The United States alone faces a shortage of nearly 200,000 people with deep analytical skills and 1.5 million managers and analysts to analyze data and make decisions on their findings[1]. These skilled workers require multiple years of mathematical training and programming experience, as well as the ability to ask targeted business questions and use data to support their conclusions.

We seek to understand these individual contributors that parse, transform, predict, and communicate data insights within organizations. I do so with the goal of accomplishing two main objectives: (i) provide insight to organizations seeking to better support or create analytical teams, (ii) and provide prospective job seekers interested in joining analytical teams as technical contributors with market and day-to-day role insight.

© 2019 Adrian Lievano All Rights Reserved

Becoming a Data-Driven Organization:

© 2019 Adrian Lievano All Rights Reserved

References:

[1] Manyika, James, et al. “Big Data: The Next Frontier for Innovation, Competition, and Productivity.” *McKinsey Global Institute*, 2011, www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20innovation/MGI_big_data_exec_summary.ashx.

[2] Henke, Nicolaus, et al. “The Age of Analytics: Competing in a Data-Driven World.” McKinsey Global Institute, 2016, www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Analytics/Our%20Insights/The%20age%20of%20analytics%20Competing%20in%20a%20data%20driven%20world/MGI-The-Age-of-Analytics-Full-report.ashx.

[3] Crawford, Chris, et al. “2018 Kaggle ML & DS Survey.” *Kaggle*, 3 Nov. 2018, www.kaggle.com/kaggle/kaggle-survey-2018.

[4] Team, Kaggle. “2019 Kaggle ML & DS Survey.” *2019 Kaggle ML & DS Survey*, 2019, www.kaggle.com/c/kaggle-survey-2019/.

[5] Team, Kaggle. “The State of ML and Data Science 2017.” *Kaggle*, 2017, www.kaggle.com/surveys/2017.

[6] Ransbotham, Sam, et al. “The Talent Dividend.” MIT Sloan Management Review, 2015, https://sloanreview.mit.edu/projects/analytics-talent-dividend/

[7] Crowdflower, Inc. “2016 Data Science Report.” Crowdflower, 2016, https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

[8] Pandey, Parul. “Geek Girls Rising : Myth or Reality!” *Kaggle*, Kaggle, 18 Nov. 2019, www.kaggle.com/parulpandey/geek-girls-rising-myth-or-reality/data?utm_medium=email&utm_source=intercom&utm_campaign=kaggle-survey-2019.

[9] Amin. “Student Community in Kaggle.” *Kaggle*, Kaggle, 26 Nov. 2019, www.kaggle.com/amiiiney/student-community-in-kaggle/comments.

[10] Bean, Randy, et al. “Companies Are Failing in Their Efforts to Become Data-Driven.” Harvard Business Review. Feb. 2019, https://hbr.org/2019/02/companies-are-failing-in-their-efforts-to-become-data-driven

© 2019 Adrian Lievano All Rights Reserved

[11] Dallemule, Leandro, et al. "What's Your Data Strategy." Harvard Business Review. May, 2017. https://hbr.org/2017/05/whats-your-data-strategy

[12] Berinato, Scott. "Data Science and the Art of Persuasion." Harvard Business Review. Feb, 2019. https://hbr.org/2019/01/data-science-and-the-art-of-persuasion

[13] Davenport, Tom, et al. "Data Not Leading to Insights? Culture Might be to Blame." Wall Street Journal. Sep, 2019. https://deloitte.wsj.com/cmo/2019/09/29/data-not-leading-to-insights-culture-may-be-to-blame/

[14] Rogati, Monica. "The AI Hierarchy of Needs." Hackernoon. June, 2017. https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007

[15] https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/why-data-culture-matters

Code Used for Analyzing Data and Creating Visualizations:

[1] Lievano, Adrian. "Adrianlievano/kaggle_data_science_2018_survey." *GitHub*, 2019, github.com/adrianlievano/kaggle_data_science_2018_survey.

© 2019 Adrian Lievano All Rights Reserved

Supporting Figures & Notes:

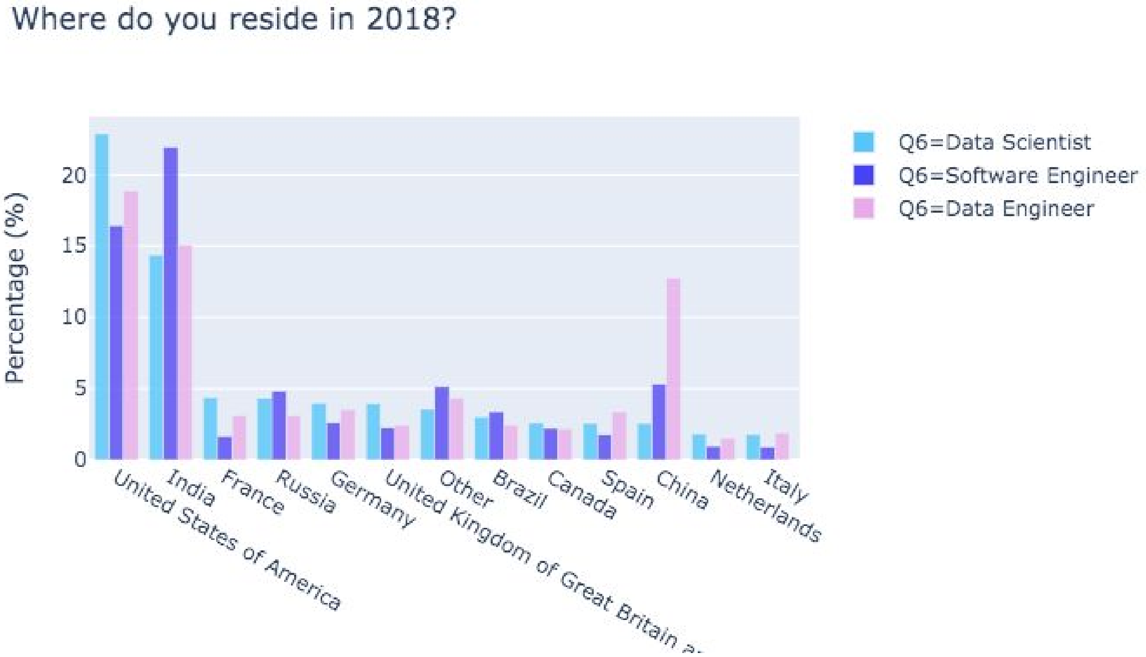

Exhibit X: Geographic concentration in percentage of self-reported data scientists, software engineers, and data engineers from the 2018 Annual Kaggle Machine Learning and Data Science Survey

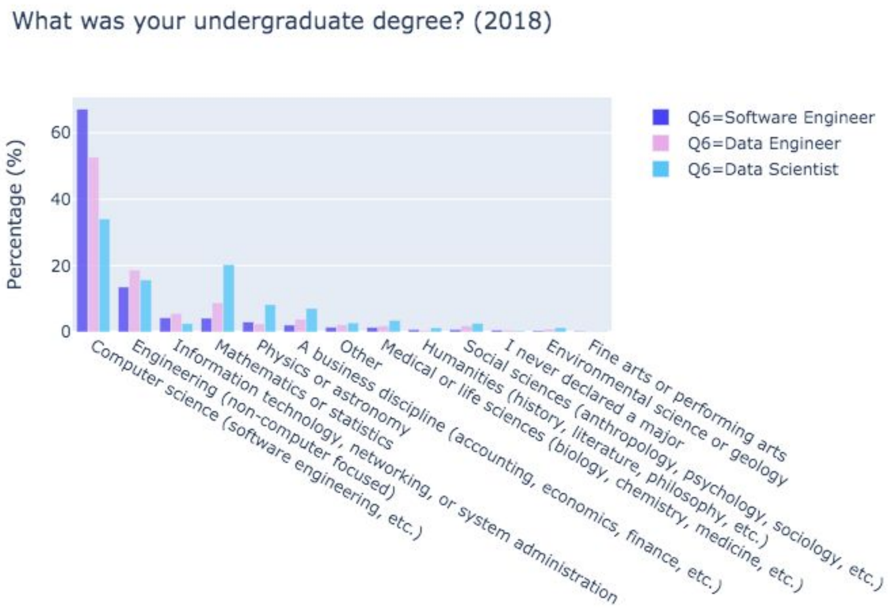

Exhibit X: Undergraduate degree breakdown of self-reported data scientists, software engineers, and data engineers from the 2018 Annual Kaggle Machine Learning and Data Science Survey.

© 2019 Adrian Lievano All Rights Reserved

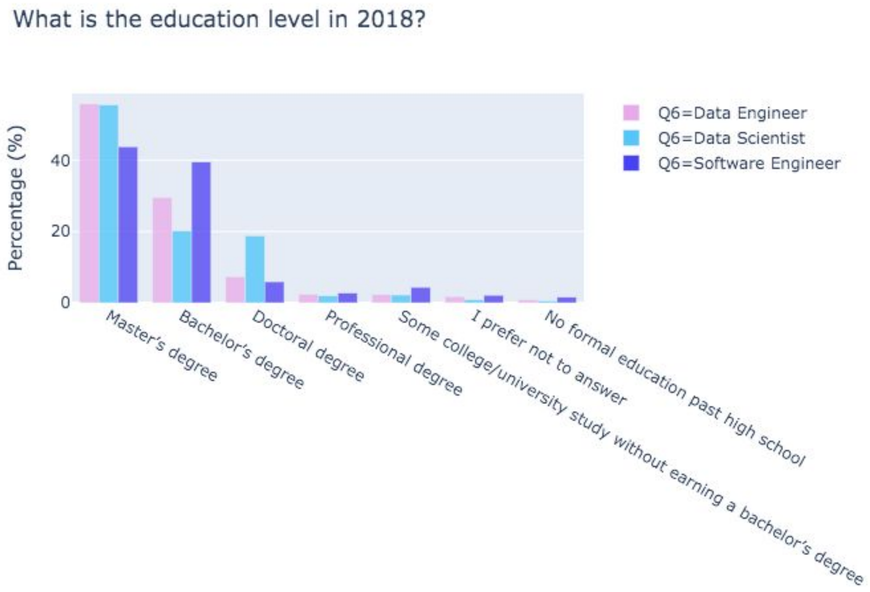

Exhibit X: Level of education for self-reported data scientists, software engineers, and data engineers from the 2018 Annual Kaggle Machine Learning and Data Science Survey.

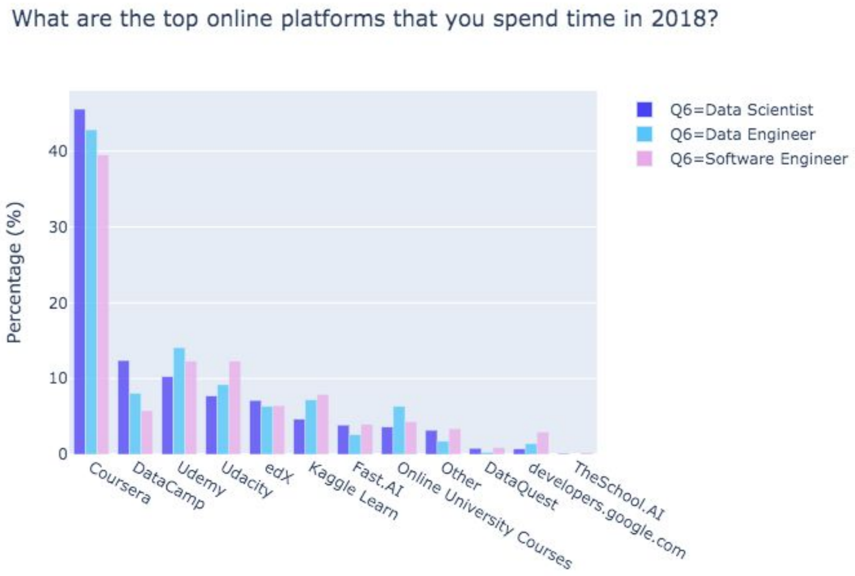

Exhibit X: Most popular MOOCs based on total time spent self-reported by data scientists, software engineers, and data engineers from the 2018 Annual Kaggle Machine Learning and Data Science Survey.

© 2019 Adrian Lievano All Rights Reserved

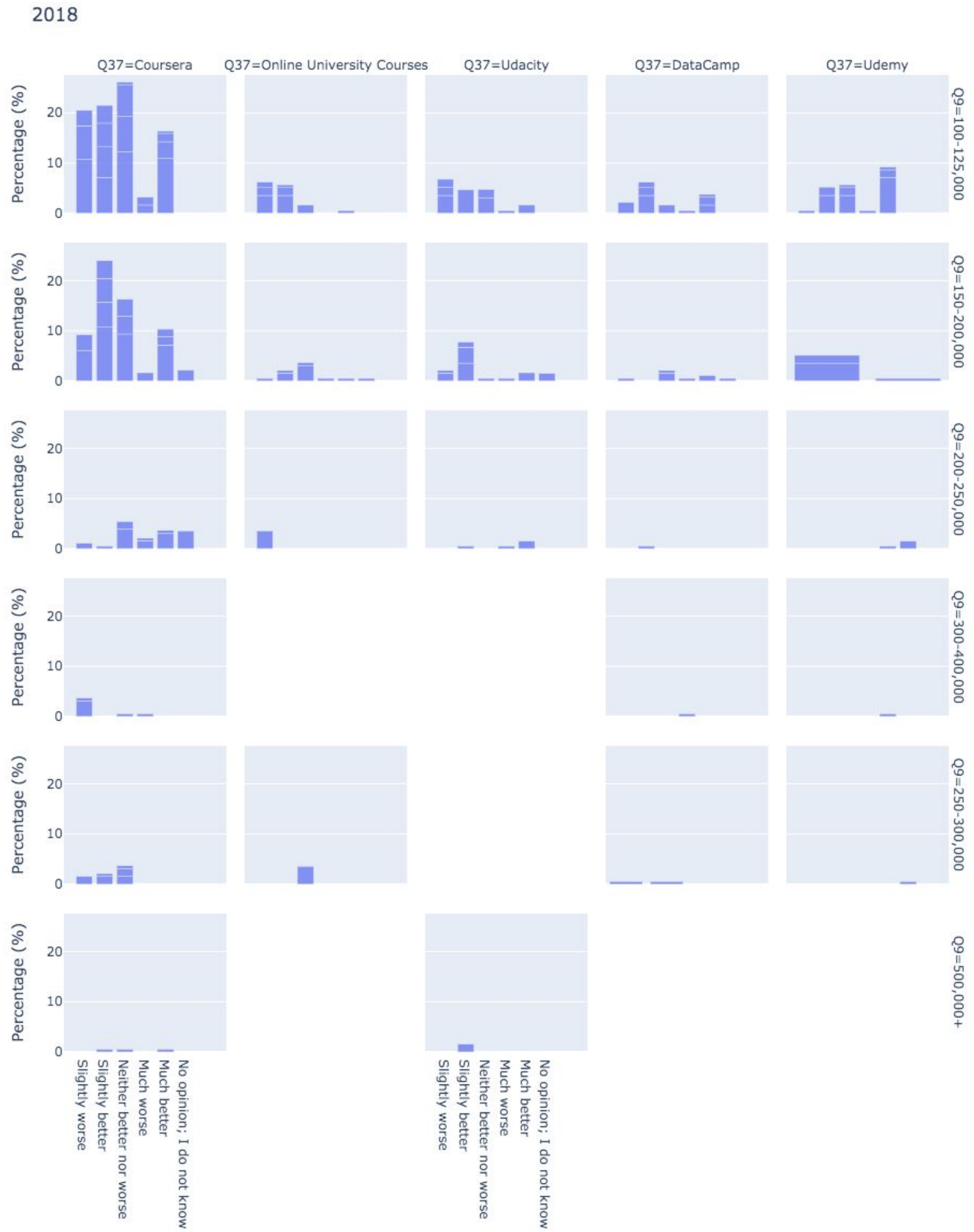

Exhibit X: Top-5 most popular Massive Open Online Courses (MOOCs) compared to traditional brick & mortar institutions. 'Much Better' indicates that a MOOC is 'Much Better' than a traditional education. This ranking includes responses from data scientists, data engineers, and software engineers separated by annual self-reported salary. Blank squares indicate no data was available.

© 2019 Adrian Lievano All Rights Reserved