# MOOD

# Massive Open Online Dataviz

**Derin Guzel, Dylan Levine, Saur Vasil, Rom Villarica**

**Advisor: Dr. Ryan Baker**

University of Pennsylvania School of Engineering

Computer Science Senior Design Project

## Abstract

Penn is part of a consortium of higher-learning institutions that work together to offer MOOCs (Massive Open Online Courses) as a way to expand their educational reach to students across the world. These MOOCs produce a large volume of data regarding student performance and online class proficiencies that would be of great interest to educators and course designers; however, the existing analytics and visualizations to derive these insights are not adequate for Penn's research and development needs. To remedy this, we created MOOD (Massive Open Online Dataviz), a specialized visualization-analytics platform for MOOCs that prioritizes ease-of-use when finding and interpreting insights. It provides course designers with a toolset to compare and contrast courses across a variety of different dimensions as well as the ability to drill down into any individual course and observe trends across time.

## Motivation

Penn, as well as a consortium of other institutions (Michigan, Duke, etc.) have come together to offer MOOCs (Massive Open Online Courses) as a way to expand their educational reach students across the world that may not have access to these resources otherwise. As one can imagine, there are massive amounts of interesting data about classes, student performance, and online class proficiencies that can be extracted; however, their form of analytics and visualizations to derive this data is not fully adequate to the consortium's research and development needs. MOOC course designers (specifically those utilizing Coursera, currently the largest MOOC platform catering to University clientele) require a comprehensive analytics tool that is simple and intuitive to use to better understand the large amounts of data about their University's courses and ultimately improve their course offerings.

The only current solution available is Coursera's offering, which is a few screens of analytics panels (that are not included in this report for data privacy reasons). These panels provide graphs of demographic information (education, age, gender, etc.) that are self-reported, as well as a basic graphs showing learners' progress over time (how many complete which modules within a course). This solution is very simple but presents only surface level insights that do not help course designers improve their courses. In addition, the data is presented in a disjoint manner, with different analytic panels found on screens that are unrelated and not named intuitively.

Throughout the course of this project, we interviewed members of the MOOC community at Penn in order to discern which factors were most important to consider in building our final product. After several rounds of interviews, we were able to figure out which factors our end users valued most and synthesize these into the below requirements:

- Capable of analyzing multiple courses at the same time such that course designers can draw parallels and make comparisons.
- Intuitive and user friendly.
- Dives into demographic as well as course data.
- Provides sentiment analysis on written feedback from the learners.

By combining data between Coursera's offerings, both within and across courses and course editions, we were able to create an app that pulled out useful insights so as to give course designers a better idea of:

1. Weak points within courses.
2. Who is actually taking these courses.
3. Students' thoughts on the courses. Combined insights (like showing that more people under the age of 21 like Wharton Online offerings) will give us a tangible advantage over Coursera and make our target users very likely to actually use our product.

## Need Finding

We had the following meetings with various stakeholders to identify their needs.

### Initial Meeting with Prof. Baker

We met with Prof. Baker and he loosely described the current state of affairs regarding analysis behind MOOC data at Penn and other schools within the consortium. Initial impressions necessitated further explorations within stakeholders to define which exact user segment we wanted to address (professors or course designers) to provide a more targeted analysis. We used this meeting to pivot to potential users and source more concrete needs.

### Wharton Online/Jessica Morris (Penn OLI)

We met with course designers at Wharton Online and Jessica Morris at Penn OLI, who took us through a deep dive of Coursera's analytic capabilities and how they used that data to improve courses within their ecosystem. Key takeaways included that Coursera provided a very broad range of data that was accessible to administrators, but this data was unorganized with limited insights (ex. basic demographic information with no link to other course performance data). We also learned that Coursera provides "cleansed" data for more advanced analytics that researchers can utilize.

Wharton Online noted that the teaching assistants are in charge of curating the course material after its initial offering, almost acting as course designers. Therefore, Wharton Online urged us to focus on course designers as the core audience. Jessica echoed this sentiment, stating that course designers would be a key stakeholder in using the application for improving their daily workflows.

We persisted from both groups on what specifically they would crave from a new application to fix the status quo. When persisting on what new data they would like, we ran into the issue of our users not knowing what could be extracted from the data that existed. This showed that our stakeholders were a non-technical group, and had essentially settled for whatever Coursera was offering..

Our next immediate insight from this meeting was a discovery that the discussion boards were a huge opportunity for insight. Currently, Coursera provides a rating system for courses and modules, but this is either left unanswered or on a 5-star scale. The interview yielded the insight that only students who actually finish something were likely to rate the course at the end, whereas people with problems either quit at a specific module or go to the discussion board to vent. This gave us the idea to perform sentiment analysis on the board posts as an augmentation of the current course rating system.

### Rebecca Stein/Jessica Morris (Penn OLI)

Rebecca Stein, a course instructor and the executive director of Penn OLI, reiterated the importance of demographics and mentioned that as an instructor she rarely looks at Coursera analytics to update the course. This cemented our focus on the course designer user segment, which necessitated a subsequent meeting with Prof. Baker to confirm our new direction.

Jessica Morris then showed us a Coursera analytics dashboard. We conducted a user study by asking Jessica to explore the application as she normally would, finding specific data queries on the dashboard. She said it was a work in progress with subpar design in terms of user friendliness (lack of meaningful visualizations, convoluted user interface). It was clear that this was the case, as she was unable to quickly find any insights, and

she used the application frequently - this made us question how users with less experience with the application would be able to glean key insights at all, and led us to the concept for our first prototype.

### Follow-up Meeting with Prof. Baker

After we presented case our for focusing on course designers (OLI) as the key user segment, Prof. Baker agreed with the decided direction of the project. We spent the rest of the meeting discussing ideal design practices and technical implementation specifics to help with beginning the product design pipeline.

## Technical Approach

### Front-end

We began our design of the front-end with a whiteboarding session that factored in the information we had gathered from our user interviews in order to come up with a rough storyboard of how we wanted the final product to look. After some discussion, we came up with a design concept that we felt would satisfy all our users' requirements: a simple one-screen dashboard that took in data, and then generated visualizations for the most-common insights our end users were interested in.

We decided to build the app using Java, as we wanted to ensure cross-platform compatibility to facilitate greater adoption. Furthermore, we were interested by the prospect of developing with JavaFX, a library similar to Java Swing in that it allows for creation of UI elements, but one that is far more extensible. We employed this library along with the ControlsFX extension to build out the components of our UI and used CSS to style them.

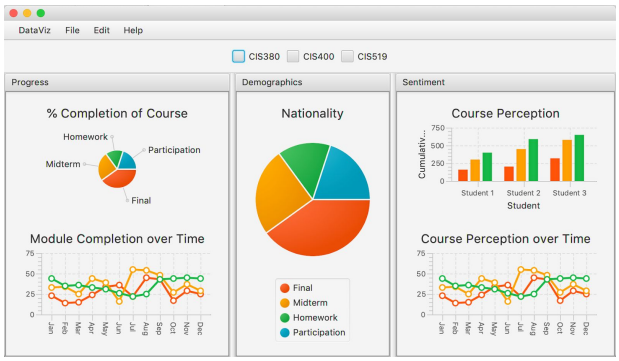



Figure 1. Our first MVP

Our first MVP, depicted above, met the initial set of requirements we had gathered. After completing it, we met with our advisor and our end users to demo what we had done and get feedback. Through this process, we uncovered another set of requirements (to be discussed in the next section), and iterated on our previous prototype to come up with one that could not only provide general insights through visualizations, but also facilitate drilling down into individual elements of each visualization in order to come up with more-specific insights. We accomplished this through designing modular, interactive UI classes that could take in datasets and produce charts while also enabling users to click on certain elements / hover over others to expose new information. This resulted in the second and final iteration of our product, shown below.

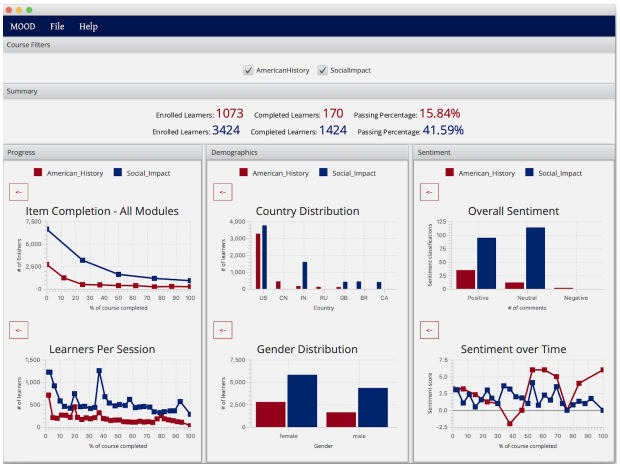



Figure 2. Our final product

Our final product is a Java dashboard that can be installed on any desktop, seamlessly integrating SQL and Python in order to present the metrics mentioned by our stakeholders in a visually-appealing, simple format. Speaking specifically with regard to the design, we drew heavy inspiration from Penn's style guide (available freely online) for the fonts and colors we incorporated into our CSS stylesheets.

### Back-end

We first started our analysis on the data received in CSV format using a python module called querycsv.py that operates on Sqlite. This allowed us to get basic queries done very easily and quickly. However, as the queries we wanted to perform got more complex, we ran into several

issues such as failure to support complex joins, casting and timestamps. We concluded that SQLite and querycsv.py are not robust enough for the purposes of this project.

The next solution we considered was setting up access to Penn's data warehouse and directly running SQL queries on the primary data. Due to security reasons, we unfortunately didn't get access to the data warehouse. Therefore, in the end we decided to construct our own database instance that interfaces our application. We chose PostgreSQL, a relational database that is recommended by Coursera, to this end. PostgreSQL is free and runs locally. It's also cited as a clone to Amazon Redshift -- the platform that Coursera uses. We believe that this choice of database and its compatibility with Coursera's platform will allow our users to easily shift to a more scalable approach in the future.

Having conducted several need finding meetings with our stakeholders such as Penn OLI, we identified two key areas of analytics for course progress data. The first set of analytics include course progress data whereby we provide the following insights:

- Drop rates across modules with drill down into specific items within a module.
- Average number of attempts per item as well as average final grade for graded items.
- Number of enrolled students in each session of the course.
- Course completion rates.

The second set of analytics are focused on the demographics of the cohorts taking the classes. These insights help our stakeholders better curate their marketing efforts and increase their market share. We provide the following insights:

- Gender breakdown with drill down into passing percentages of the students.
- Country breakdown of the students with drill down into passing percentages and state/region breakdown if applicable.

Upon consulting our advisor, Dr. Baker, and discovering that Coursera offered no current tools for analyzing user sentiment of courses based on feedback comments, we decided to incorporate the feature into our application in addition to the course data analytics. After researching a number of diverse natural language processing tools, we decided to use the AFINN lexicon, an open-source dictionary containing over 3,300+ words with a polarity score associated with each word. The lexicon has a wrapper library in Python called **afinn**, which is what we used to conduct the analysis.

| Course Title | Number of Comments | Relevant Comments* | % Yield | Correctly Evaluated Comments | % Yield |
|---|---|---|---|---|---|
| American History | 50 | 42 | 84% | 36 | 86% |
| Microeconomics | 341 | 258 | 76% | 238 | 92% |
| Social Impact | 210 | 124 | 59% | 102 | 82% |

Table 1. Sentiment Analysis Evaluation

*This reflects the number of comments that directly relate to the quality of the course

The NLP analysis begin as an isolated Python script that locally scanned CSV files containing user feedback comments for different courses and analyzed the net sentiment scores of each comment on afinn's scale of -10 to 10, graphing both the number of positive comments (those with a score > 0), neutral comments (those with a score = 0), and negative comments (those with a score < 0), as well as a graph of average sentiment scores per course session. From there, we integrated the code into our JavaFX program with the help of PyDev and Java's Runtime and ProcessBuilder classes, which allowed us to use Eclipse to run the python script in Java and retrieve the output values. Upon the introduction of a PostgreSQL database instance for storing our course data, we re-integrated the code with the help of Psycopg, a PostgreSQL database adapter for Python.

In an attempt to increase the accuracy of the NLP analysis, we eventually integrated a Google translate (googletrans) and spell checker (pyspellchecker) library in order to ensure that every word parsed by the analyzer script would be recognizable to the lexicon, which could only detect English words and did not automatically correct for spelling mistakes. Introducing these libraries required us to slightly modify the flow of the application such that translation and spell checking of course comments could be periodically run as a form of database maintenance, rather than running them each time a course is selected for analysis, which oftentimes took longer than what we considered to be practical. Because of the limits of the dictionary we were using to conduct spell checking, we additionally had to add a number of niche acronyms and phrases to the lexicon (e.g. "MOOC", "Coursera", etc.) such that these strings would not be falsely corrected.

## Evaluation

### Front-end

We had several meetings with our advisor, Dr. Baker, as well as with the OLI team to showcase the intermediate and final iterations of MOOD. Both groups expressed high praise for the intuitive UI, as well as appreciating the use of NLP for sentiment analysis. The OLI team especially appreciated the application's ability to easily compare arbitrarily many courses at once, as well as the inclusion of intuitive drill down functionality with many layers of insights for each category (e.g. the number of passing users per state in the US, etc.) - all of which Coursera's standard analytics tool does not currently provide. Additionally, the OLI team commented on the professionalism of the look and feel and commended our use of Penn's style guide in creating the aesthetics of our final design.

Potential improvements they mentioned primarily revolved around clearing up some of the descriptions on the dashboard so as to be less in-line with what Coursera provided and more human-readable, as well as making an indicator to clarify which level of drill-down the user was currently at when looking at a particular chart. We will be certain to take these into account in the case that we pass this project down to future teams.

### Backend

We created mock data to test the accuracy of our scripts. This helped us debug more complex and analytically involved queries such as the average number of attempts per item in the course. Furthermore, we were able to compare and contrast the queries across different courses. By doing so, we encountered more edge cases. Our queries are consistent across different courses, account for multiple attempts per user, maintain consistency in number of items and successfully orders modules and items.

Upon reviewing the online feedback for our 3 courses, there are some metrics we have calculated to determine the accuracy and reliability of our current algorithm for sentiment analysis. These metrics are summarized in the table on the previous page. Overall, of the 424 comments whose content was directly related to the quality of each course, 376 were evaluated correctly, yielding an 89% rate of sentiment accuracy.

## Discussion of findings, success in addressing user needs

We are proud that we were able to develop an application that is able to legitimately assist our end users. With MOOD, both course designers and researchers will be able to successfully gain more insights faster than the status quo (Coursera's standard analytics tool). In addition, our users will have access to unique insights that Coursera simply doesn't offer, like the ability to compare multiple course offerings over similar metrics and observing sentiment analysis over the lifetime of a course edition.

When conducting final user testing, our focus was on 1) determining the speed improvement of finding insights that could have also been found on Coursera's platform, and 2) the quality of insights that could be derived. The quality of insights was further segmented by the insights MOOD could uniquely generate, and the insights that were more clearly stated than the status quo.

The result was overall positive. Both user segments were able to locate more relevant information faster, and in a clearer manner than Coursera's inbuilt tool. Specific

feedback from our final testing focused primarily on methods of clarifying what specific metrics were referring to, such as differentiating between attempts (number of times an assignment was performed) vs completion (number of times an assignment was finished). More of this was discussed in the section above (front-end evaluation).

For further steps, our Dr. Baker has recommended we look towards integrating our system with Penn's backend interface, in order for our stakeholders to actually use the final project. On a technical front, this means removing our backend integrations and replacing them with the Coursera data pipeline, which is in an identical format. We are looking towards perhaps integrating this during the summer.

We may also work with a team next year to continue our progress where we left off, in order to ensure that the project continues to receive continuous updates and functionality additions.

## Ethical considerations and societal impact

Our project began with an in-depth exploration of the ethical consideration we would need to address, particularly because we would be dealing with massive amounts of user data. This user data took the form of Coursera collected metrics and personal details (location, education, age, race, gender, etc.)

In evaluating our ethical considerations, we began with contemplating who the target users of our final application would be. Upon completing our need finding process in our initial design stages, we settled on focusing on the Penn OLI (course designers) and Education School (researchers) as our user segments.

With both groups, the people actually using MOOD may or may not have clearance to actually view the user data from Coursera. This meant that, if we wanted to present a complete solution, we would need to make our application hides this raw user data, and only focuses on the insights from the data. We also needed to focus on ensuring that none of our insights could result in users deriving personally identifiable information regarding Coursera users.

Both of these elements were constantly considered during our development process. We were also able to verify that our target users (both of these Penn departments) were not able to access any Coursera user information while using our app.

## Business plan

### Value proposition

As outlined previously, our value proposition stems from the inadequacy of currently available solutions such as the analytics platform of Coursera. The present tools our stakeholders rely on lack user friendliness in the front-end and robust analytic insights in the back-end. More specifically, course designers and researchers suffer from convoluted user interfaces where the data is presented in a disjointed manner. Furthermore, given the insights provided are not very vigorous, our stakeholders fail to draw actionable conclusions from the data. We believe that the user centric development cycle we adopted while building MOOD addresses both of these concerns. Having incorporated the ability to compare courses, analyze demographic as well as course progress data while maintaining a user-friendly UI, we are confident that MOOD will meet the market need.

### Stakeholders

We categorize our stakeholders who are education providers into two groups based on their motivations for data analysis. The first group is composed of course designers who wish to identify the problems with course content in an attempt to improve the students' learning experience and eventually increase course completion rates. Individuals in this group are interested in insights such as but not limited to module drop rates, average number of attempts per item and sentiment analysis regarding course content.

The second group is composed of individuals who are more business-oriented in that their main purpose is to increase market share of the courses. To this end, they want to penetrate new market segments and maximize customer lifetime value by increasing customer retention. By the nature of their goals, this group is more interested in demographic data. They want to get a better sense of who is taking the course in order to curate their

marketing efforts and follow a more customer centric approach.

Although the two groups differ in their end goal, they share other characteristics and suffer from similar issues. First, neither of the groups is tech-savvy. A viable solution should take this into account. The final product should not only maintain a user interface that is simple and straightforward while presenting the insights but also seek to supply ease of use at other stages in the data analysis cycle such as the uploading of raw data. In addition to this, the quality of the course carries a dual purpose for both groups. Even though it is the primary concern for the first group, the second more business-oriented group may benefit from it as well. It is possible to increase market penetration and customer retention by offering courses that provide an excellent learning experience. Therefore, MOOD's course analytics data as well as the sentiment analysis serve the needs of both of our stakeholders.

We were lucky enough to work in collaboration with our stakeholders on campus. We collaborated with the Graduate School of Education (GSE) through our advisor Dr. Baker. We categorize the GSE into the first group of our stakeholders who are more concerned with course content. Penn Online Learning Initiative (OLI) is another stakeholder we got to interact with on the ground. Penn OLI aims to increase Penn's online course reach by centralizing the online classes provided by the eight different schools at Penn. Given the nature of their purpose, they are both interested in course content and demographic data.

## A Market Opportunity to Serve Education Providers

The market for online courses is growing rapidly. In the near future, education providers, such as Penn and other members of the consortium like Duke, will find themselves in an increasingly competitive landscape filled with opportunities and threats. Therefore, it is critical for education providers to remain up to date with the latest trends in the online courses market.

A recent survey of 234 education providers conducted by Tagoras, a consulting firm specialized in e-learning, show that education providers are primarily concerned with increasing their efforts to gather and analyze data to aid in creating new products, improving the existing ones and demonstrating the effectiveness of the e-learning experience offered.

Despite this need for robust data analysis tools, even the largest MOOC provider Coursera with its customer base of over 33 million fails to meet the demands of education providers. Therefore, we believe that MOOD, a specialized visualization-analytics platform for MOOCs that prioritizes ease-of-use when finding and interpreting in-sights, can successfully fill this void in the market.

Our potential customer base for this product consists primarily of course analysts and course designers. Since this customer base is by nature affiliated with the organizations sponsoring these courses, we can obtain a rough estimation of the size of the potential market by assessing these organizations: specifically, those that use Coursera as their platform.

From visiting the Coursera website, we observe that they have 186 partners across 43 countries, offering a total of 3,541 courses. These partners range from reputable universities such as Penn, Stanford, and Yale to large commercial firms such as Google, the Boston Consulting Group, and Goldman Sachs. Assuming a 70% / 30% split, reasonable given Coursera's greater focus on universities, that leaves us with 130 universities and 56 commercial firms.

Of the current market offerings, MOOD is most similar to business intelligence (BI) platforms such as Tableau and Microsoft Power BI. From a Forbes article on BI platform adoption rates, we note that commercial firms have an average adoption of roughly 40% while higher education has an average BI adoption of roughly 25%. Using these as benchmarks, we project 0.4 x 56 + 0.25 x 130 = 55 clients in Year 1. We note that university adoption would likely be higher given MOOD's greater focus on the education space.

## Competition

Our primary competitor in this space would be the default tools offered by the course provider itself: in this case, Coursera's built-in analytics. However, from the preliminary user interviews, we found that this platform was sorely-lacking in anything but the most basic of insights: additionally, Coursera's platform was based on

breadth of insights rather than depth of insights, resulting in pages and pages of charts and graphs that led to an arduous and uninformative user experience. In order to differentiate ourselves from the existing solution, we adopted the opposite approach, choosing instead to focus on depth of insights. Through our interactive single-screen dashboard featuring the ability to drill down into almost every element of every chart, we greatly reduced information inundation and created a specially-tailored user experience for course analysts that presents the most important insights in a more-concise manner than Coursera's built-in platform.

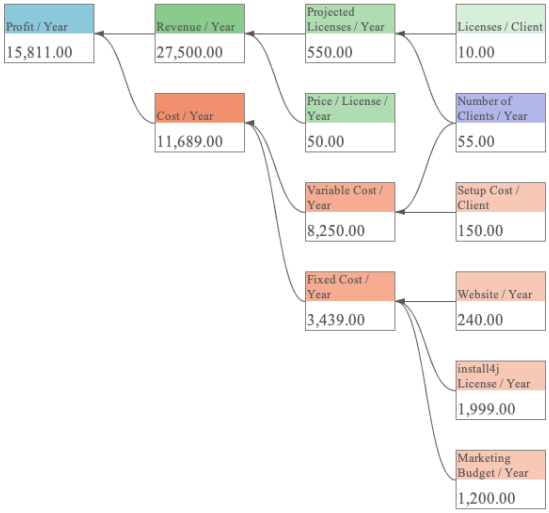
### Revenue Model and Cost Analysis

With regards to our pricing strategy, we began by attempting to determine the value MOOD adds to organizations through improved analytics. The main roadblock we encountered was that in this case, since MOOD provides insights that Coursera's built-in analytics platform does not support (such as comparing different courses on a variety of dimensions), quantifying value added by examining improvements to the existing workflow was almost impossible (made more so by the fact that we did not have precise data on the existing workflow to begin with).

Instead, since MOOD is essentially a desktop analytics platform specialized for course data, we opted to look at similar products in the desktop analytics market in order to construct a benchmark from which we could determine a reasonable price. Much like our development process for the actual application, we started with a generalization, then drilled down: our first step was to look at Tableau, a data-visualization product that enables users to create beautiful visualizations of any dataset, though non-interactive and focused more on general analytics than the specific needs presented by our potential customer base. Their subscription pricing model for businesses ($70 per desktop per year) seemed like an excellent starting point: since we are primarily B2B and aimed at larger organizations with a much-narrower focus, we thought that $50 per desktop per year would be suitable. This is significantly cheaper than most other products of this kind: we hope to incentivize greater adoption through competing on both features and price.

Cost-wise, our fixed costs would consist of expenses related to our website (using Wix Business Professional), our Java installer distribution license (install4j, an application that builds out Mac/Windows installers from .jar files), and a very small marketing budget for Google AdWords ($100 / month). Since any additional work required to onboard new clients would be a single instance per client, we will assume variable costs to be on a per-client basis rather than per-unit. We estimate that due to the modularity of MOOD's design, any adjustment costs needed to customize the application for a new client would be less than 4 hours of work, valued at roughly $150. Furthermore, we estimate (from observing Penn's situation) that each client would require around 10 licenses for their course analytics team.

Hence, our projected revenue model for Year 1 would be as follows. All numbers not previously discussed are sourced directly from the relevant companies' websites.




Assuming that we achieve this level of adoption, in subsequent years the variable cost per client will drop to 0 (as we will have already done the work of onboarding them), substantially increasing our revenue stream and thereby our profit margin.

## Conclusions

This application was built off of the standard Coursera endpoint that every partner university has access to, our application can also be extended to other institutions in

the Coursera Consortium. This means that, eventually, course designers and researchers from all over the world can use our tool to gain unique insights that previously would have been unavailable. These insights will be used to market course offerings better, tailor specific courses to their newly understood target market, and improve courses overall based on new ways to understand course feedback.

We look forward to seeing how MOOD can change Massive Open Online Courses for the better.

## References


Columbus, Louis. "The State Of Business Intelligence, 2018." *Forbes*, Forbes Magazine, 8 June 2018, www.forbes.com/sites/louiscolumbus/2018/06/08/the-state-of-business-intelligence-2018/#6c4344e27828.

"The Market for Online Courses – 2019 and Beyond." *Learning Revolution*, 2 Jan. 2019, www.learningrevolution.net/market-for-online-courses-2019/