Data science skills for the built environment: Lessons learned from a massive open online Python course for construction, architecture, and engineering

Clayton Miller1,* and Charlene Tan1

1Department of the Built Environment, College of Design and Engineering, National University of Singapore (NUS), Singapore

Abstract. It’s not just the models, techniques, or technologies that improve building performance; the digital skills of built environment professionals also play a significant part. The deluge of data from buildings, intelligent systems, and simulation tools is well-documented, and like other domains, building design, construction, and operations professionals are keen to learn skills like Python scripting that are common to the data science communities. This paper analyzes a massive open online course on the edX platform called Data Science for Construction, Architecture, and Engineering. This course was launched in April 2020, and it combines building science concepts with beginner-level data science skills, such as using Python and the essential libraries of Pandas, Sci-kit Learn, and Seaborn. This paper presents an analysis of the demographics and geographic data from 18,600 participants and survey results from 126 out of 1,561 verified course users. The survey focused on the experience of course participants and suggestions for improvement. This information can aid other data science educators in developing content to better educate built environment professionals.

Introduction

The building industry is exploding with data sources that impact the built environment’s energy performance and the occupants’ health and well-being. Spreadsheets just don’t cut it anymore as the sole analytics tool for professionals in this field. Participating in mainstream data science courses might provide skills such as programming and statistics; however, the applied context to buildings, which is the most critical part for beginners, is missing.

This paper analyzes an online course focusing on developing data science skills for professionals, specifically in the built environment sector. It targets architects, engineers, and construction and facilities managers with little programming experience. Data science skills are introduced in the context of the building life cycle phases. Participants use large, open data sets from building design, construction, and operations to learn and practice data science techniques. The course was designed to add new tools and skills to supplement spreadsheets. Major technical topics include data loading, processing, visualization, and essential machine learning using the Python programming language, the Pandas data analytics and sci-kit learn machine learning libraries and the web-based Colaboratory environment. In addition, the

*e-mail: clayton@nus.edu.sg
course provides numerous learning paths for various built environment-related tasks to facilitate further growth.

**Built Environment Data Science Education**

To put the analysis of this course in context, the background of using data science as a tool in the digitization of the built environment is analyzed. A recent study showed that research, social, economic, and technological factors were the primary drivers influencing successful construction digitization [1]. Their study highlighted the paramount role of the social factor, which exerts a widespread influence across construction processes. Achieving true digitization necessitates the comprehensive integration of all involved parties. Given the multi-party nature of construction projects, the onboarding of every participant becomes a crucial prerequisite. However, executing this integration poses significant challenges. Each involved party faces unique implementation hurdles, compounded by the inherent complexity of construction projects, including project intricacy, uncertainty, fragmented supply chains, short-term perspectives, and cultural barriers [2]. Consequently, the digitization of this sector confronts numerous hurdles.

Additionally, [3] discovered that despite having the necessary data, operational capabilities, technical support, and a conducive culture, only 7.1% of construction organizations had adopted big data and predictive analytics by 2020. These analytics were potentially applicable in cost estimation, delay and consumption prediction, and energy management. Despite adequate resources, the study’s revelation of low implementation rates led Ngo to speculate that the scarcity of technologically skilled professionals within the sector might be the primary impediment. This emphasizes the critical need to adequately train and equip Built Environment professionals to navigate and capitalize on the digital revolution.

This paper explores the deployment of an online course designed to address these challenges through practical training in using the modern, high-level data Python programming language and its various libraries. This effort covers the course’s structure and content, the participants’ demographic information extracted from the edX platform, and the survey results given to the verified track participants to explore their feedback and interest in future work.

**Course Structure**

The course’s structure focuses on the various phases of the built environment and the application of Python-driven data science skills in each life cycle phase. Table 1 outlines the course sections in detail, including the skills, phase targeted, and real-world applications. The introduction overviews key Python concepts and the motivating factors for building industry professionals to learn to code. The SDE4 Net-Zero Energy building example is introduced as a scenario that uses various data science-related technologies in its design, construction, and operations [4].

Section 2 then provides the foundational functions of the Pandas library, which are demonstrated in the integrated design process by processing data from parametric EnergyPlus models. Future learning path examples are introduced for the Design Phase, including building information modeling (BIM) using Revit or Rhino, spatial analytics, and building performance modeling Python libraries.

The course then moves onto Section 3, where time-series analysis using Pandas functions is demonstrated in the Construction Phase by analyzing hourly IoT data from electrical energy meters from the Building Data Genome 2 Project [5]. Further examples of future learning...
paths are introduced for the construction phase, including project management, building man-
agement system (BMS) data analysis, and digital construction, such as robotic fabrication.
Section 4 then covers various statistical aggregations and visualization techniques using Pandas and the Seaborn library, demonstrated on Operations Phase occupant comfort data from the ASHRAE Thermal Comfort Database II [6]. Further examples of future learning paths are introduced for the Operations Phase, including energy auditing, IoT analysis, occup-
ant detection, and reinforcement learning.
Finally, Section 5 overviews the motivations and opportunities for using prediction in the built environment. It then demonstrates prediction, classification, and clustering using the sci-kit-learn library on electrical meters and occupant comfort data. The course concludes with suggestions for more in-depth Python, Data Science, and Statistics courses.

The course has two types of participants: Regular and Verified Track users. The regular track is for passive users who want to view and learn from all the course content for free. For Verified Track users, the course is graded through two quizzes from each section, and a certificate is provided upon successfully completing the course. To achieve the Verified Certificate, the participant must have a grade of 75% or higher. There are 10 total quizzes; the lowest two are dropped for the final score. Each quiz has ten total points available, seven of which are multiple-choice questions, and then there are points available for the answers to coding exercises given for each section.

**Participant Demographics**
Throughout four consecutive launches, the course amassed an enrollment count that sur-
passed 38,000 participants, of which around 34,000 are currently still enrolled. The most recent, largest, and longest launch was from August 2021 to the present, comprising over 18,600 participants. This section analyzes the self-reported aggregated information from the participants from the most recent launch and outlines an overview of this data set’s education levels and geographic distribution.

**Geographic Distribution**
Figure 1 shows the breakdown of learners from across various regions of the world from the latest launch of the course that has taken place since August 2021. Participants from 184 countries or regions in the world were represented, with India, the United States, and the United Kingdom providing the most self-reported learners. These three countries make up over 30% of the total learners. Other countries with significant numbers of participants include Singapore, Canada, Germany, Brazil, Türkiye, and Egypt.

**Level of Education**
Figure 2 depicts the known educational backgrounds of the course learners who shared this information. Over 75% of the participants have a Bachelor’s Degree or higher. These data reveal that the average participant is highly educated and a degree-holding learner. The re-
sults of these findings were not surprising, as literature from [7] found that the majority of MOOC students hold a higher educational status, which was speculated to be due to educated individuals having more access to MOOCs in terms of internet access, knowledge access, etc.
Table 1. Course Outline

<table>
<thead>
<tr>
<th>Section</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Section 1</td>
<td>Introduction to Course and Python Fundamentals</td>
</tr>
<tr>
<td></td>
<td>• Overview of the course, tools, and data</td>
</tr>
<tr>
<td></td>
<td>• Phases of the building industry: design, construction, operations</td>
</tr>
<tr>
<td></td>
<td>• Key Python concepts: Variables, Control Statements, Data Structures</td>
</tr>
<tr>
<td></td>
<td>• Real-World Data Science Applications Overview</td>
</tr>
<tr>
<td>Section 2</td>
<td>Introduction to the Pandas Data Analytics Library and Design Phase Application Example</td>
</tr>
<tr>
<td></td>
<td>• Design Phase of Buildings overview</td>
</tr>
<tr>
<td></td>
<td>• Pandas data analysis library introduction</td>
</tr>
<tr>
<td></td>
<td>• Basics of Pandas data structures and common functions</td>
</tr>
<tr>
<td></td>
<td>• Applications for design: EnergyPlus, Revit, Rhino, spatial analytics, generative design</td>
</tr>
<tr>
<td>Section 3</td>
<td>Pandas Analysis of Time-Series Data from IoT and Construction Phase Application Example</td>
</tr>
<tr>
<td></td>
<td>• Construction Phase of Buildings overview</td>
</tr>
<tr>
<td></td>
<td>• Time-series data and its challenges exploration</td>
</tr>
<tr>
<td></td>
<td>• Deep dive into Pandas functions specific to time-series data</td>
</tr>
<tr>
<td></td>
<td>• Applications for construction: project management, building management system (BMS) analysis, digital construction</td>
</tr>
<tr>
<td>Section 4</td>
<td>Statistics and Visualization Basics and Operations Phase Application Example</td>
</tr>
<tr>
<td></td>
<td>• Operations Phase of Buildings overview</td>
</tr>
<tr>
<td></td>
<td>• Discussion on occupant satisfaction in buildings</td>
</tr>
<tr>
<td></td>
<td>• Pandas statistical aggregations and visualization techniques</td>
</tr>
<tr>
<td></td>
<td>• Introduction to Seaborn visualization library</td>
</tr>
<tr>
<td></td>
<td>• Applications for operations: occupant satisfaction, spatial optimization, feedback collection</td>
</tr>
<tr>
<td>Section 5</td>
<td>Introduction to Machine Learning for the Built Environment</td>
</tr>
<tr>
<td></td>
<td>• High-level overview of prediction’s impact on the building industry</td>
</tr>
<tr>
<td></td>
<td>• Implementation of prediction, classification, and clustering using sci-kit learn library</td>
</tr>
<tr>
<td></td>
<td>• Application to smart meter and occupant comfort data</td>
</tr>
<tr>
<td></td>
<td>• Machine learning applications in the building industry, Kaggle competition example</td>
</tr>
<tr>
<td></td>
<td>• Suggestions for further Python, Data Science, and Statistics courses on edX</td>
</tr>
</tbody>
</table>
Figure 1. Breakdown of the geography of the participants from the latest deployment of the course.
Participant Feedback Survey

In addition to the self-reported demographic data provided by the platform, the course also included an optional survey conducted after the course, where participants were given a chance to give feedback about the course, its usefulness, and some insights on how to improve it. The total number of Verified Track users eligible to complete the survey was 1,561, and the completed survey sample size was 126. With a 95% confidence level, this survey has a margin of error of 8%.

This survey took the form of a web-based questionnaire distributed among learners participating in all course launches to evaluate the course. Employing a survey questionnaire for course evaluation is a prevalent practice that delineates various purposes for such course evaluations, including course enhancement, by identifying satisfactory elements and rectifying areas requiring improvement [8].

The questionnaire was crafted to encompass both closed and open-ended inquiries, signifying the utilization of a mixed-methods research approach. This approach, defined by [9], amalgamates traditional quantitative and qualitative research facets, forming a distinct research paradigm termed mixed methods. Among research methodologies, mixed methods studies are posited to yield the most comprehensive, balanced, and informative outcomes.

Given the global audience, an online survey was the most viable means of collecting responses, rendering physical questionnaires impractical. The survey link was shared in the concluding section of the course, and participation was entirely voluntary for learners. This paper focuses on a subset of the questions that captured quantitative ratings, which were given mainly about the quality of the course, as well as qualitative discussion questions that allowed the sharing of suggestions for improvement.

Quantitative Feedback

Figure 3 gives an overview of the quantitative questions asked of the participants. Four questions were given that requested information about the participant’s feelings about the course, whether it kept them motivated, whether they would recommend it, and how well-developed the course content was for their needs. The survey respondents would indicate their feedback on a range from 1-5, with 1 indicating a low opinion or reflection of the question and 5 indicating a high opinion.

Figure 2. Breakdown of the level of education of the participants from the latest deployment of the course
Figure 3. Outline of the survey responses from verified track participants

Qualitative Feedback

The quantitative analysis of survey responses indicated respondents’ generally positive impressions of the course. However, uncovering areas for improvement proved challenging through closed-ended questions alone. The strategic inclusion of open-ended questions for each course week significantly facilitated the identification of specific aspects requiring adjustment or retention. This approach met the typical objectives of a course evaluation survey and aligned with the overarching research goal of enhancing the course’s quality.

Course Specificity

The course’s precise title suggested involvement in the built environment sector for enrolled participants. While numerous online introductory data science courses exist, their applicability to real-life scenarios in the built environment sector remains limited. This leads to an Information Paradox, a situation with abundant information but scarce practical knowledge. Feedback emphasized the rarity of courses dedicated to data science within this sector, acknowledging the course’s unique contribution to bridging these knowledge gaps. The course content already addressed some of these gaps, amplified by the sector’s burgeoning interest, amplifying its value.

Satisfaction with Course Content and Teaching Methods

Respondents expressed interest in future courses from the developer, and suggestions for advanced topics within data science and machine learning tailored to the building industry signaled satisfaction with the current course. Most responses in key evaluation questions underscored positive perceptions of the course overall, particularly evident in the high scores for course rating and well-developed aspects.

One of the survey questions asked for information about the respondent’s interest in potential topics that could build upon the current course. Logical deductions were made to glean more profound insights into the respondents’ perceptions of data science. If a respondent selected any available options or chose ‘Others’ while specifying additional courses related to this course, it was inferred that the individual expressed interest in advancing their data science knowledge within the Built Environment sector. Conversely, if a respondent did not
select any options, regardless of providing reasons, they were considered as showing no inclination toward further learning. While the level of measurement remained unchanged, this adaptation conferred specific meaning to the responses.

**Discussion**

Much of the qualitative feedback in the participant survey can be further discussed in the context of learning from the course’s deployment and improvement. Other educators and course developers can also learn from this experience.

**Diversity of Topics Covered**

Feedback from various weeks of the course highlighted issues with misrepresented or inadequately covered data specific to the built environment sector. For instance, Section 2 of the course, which discusses the design phase, only lightly touches upon BIM model data without demonstrating its practical application in real-life scenarios. Given the sector’s heavy reliance on BIM in digitalization, learners expected deeper insights into utilizing data science with BIM datasets. The course’s failure to meet these expectations resulted in disappointment among learners, revealing a persistent knowledge gap. Addressing this could involve integrating open-sourced BIM model datasets into the curriculum for analysis, showcasing practical applications. Similarly, in Section 3, several learners expressed dissatisfaction with construction-related content, expecting more relevant and detailed data.

**Expansion of Data Science Applications**

The desire for increased avenues to apply data science extended beyond the previously mentioned questions. Notably, several responses to one of the qualitative questions echoed similar sentiments, emphasizing the widespread appeal for diversified applications. The prevalence of these requests underscores significant knowledge gaps awaiting fulfillment. Respondents’ specific examples of desired data for the course offer valuable insights into identifying these knowledge gaps. Engaging past learners active in the field by soliciting their datasets could be instrumental in enhancing the course’s relevance. Incorporating real-life datasets into the curriculum aligns the course content with real-world scenarios, potentially expediting the practical application of data science in their professional realms.

**Content Depth**

Section 5, focusing on Introduction to Machine Learning, generated considerable feedback regarding content depth. Learners expressed dissatisfaction with the video explanations’ lack of detail, which caused an abrupt surge in difficulty compared to prior weeks. Their preference leaned towards a more gradual learning curve, suggesting the possibility of splitting the section into two smaller sections. This suggests a dual narrative: learners acknowledged the dense content yet remained eager to delve deeper into machine learning, even advocating for an extended time frame to grasp the concepts thoroughly.

Conversely, another subset of learners felt the week’s content was merely introductory, advocating for expanded coverage in a subsequent course. Despite the preceding weeks’ beneficial data analysis, the expressed interest in machine learning implies a professional inclination toward automated data utilization rather than mere analysis. Here, the course deals with two perceptions of the content – either too heavy or insufficient. For the former,
the course developer can split the content into two weeks or provide more support through discussion boards or additional online resources. For the latter, an audited and approved follow-up course could be recommended to learners interested in learning more.

Conclusion

This paper conducted an exploratory analysis of data from a course evaluation survey. The analysis revealed a pronounced interest in applying data science within the built environment context, yet it unveiled substantial knowledge gaps hindering built environment professionals from effectively utilizing this data. The paper successfully identified some of these knowledge gaps, providing recommendations for their mitigation.

Limitations

The survey’s primary limitation lies in its original intent as a course evaluation tool rather than a research instrument. Consequently, the need for data conversion arose to render questions analyzable. Despite logical conversion attempts, this method’s absence from existing literature raises concerns about its impact on the results. Designing a research-oriented survey aligned with specific objectives could yield more precise and meaningful data, avoiding such conversions. The subjectivity inherent in survey responses is another critical limitation. Responses were wholly reliant on individual perceptions and understanding of the questions, potentially leading to inaccuracies due to human error, faulty recollections, or biases.

Moreover, the survey’s voluntary nature introduces voluntary response bias, wherein individuals with strong opinions are overrepresented. This aligns with findings suggesting dissatisfied individuals are more inclined to provide feedback, potentially skewing the data. The concept covered by [10] related to non-response bias is relevant here, as the absence of responses might suggest contentment, skewing assumptions toward the course’s satisfaction. However, despite these limitations, the research findings effectively achieved the study objectives, shedding light on ways to enhance the course and bridge the gap between data science and the built environment sector.

Areas of Further Development

The existing literature deficit in applying data science within the built environment context necessitates further investigation to expedite the adoption of digitized processes. Future research avenues include:

- Assessing the industry’s awareness of data science benefits and its receptiveness to coding education, including the utilization of emerging generative coding tools.
- Identifying challenges impeding built environment professionals from coding and discovering strategies to overcome them and what type of tasks and workflows are most applicable to be impacted using data science.
- Exploring the motivation driving data science adoption in the building industry, understanding challenges motivating its usage to aid others facing similar issues.
- Investigating the benefits of applying data science for the built environment, targeting participants who completed the course or practicing professionals to document advantages that could boost the adoption of data science.
- Integration of coding automation tools that utilize large amounts of code examples and leverage Large Language Models (LLM), such as ChatGPT [11].
• Utilization of machine learning competition scenarios that help participants quantify the development of their skills through the scoring of production submissions for accuracy [12–14].
• An expansion of the use of open data sets from a wider range of sources [15] from the built environment including images, video, and text [16]
• Better capture the human aspects of the built environment through integrating data from wearables, smart devices, and occupant interaction [17, 18].
• Alternatively, within the MOOC landscape, research opportunities exist in studying learner engagement across certified versus non-certified learners and exploring demographic influences on learning experiences.

The paper aimed to uncover knowledge gaps among those interested in applying data science in the built environment context, evidenced by the enrollment of over 38,000 participants. The literature review exposed inadequacies in the digital skill sets necessary to harness industry data effectively. Feedback predominantly comprised recommendations, emphasizing the need for real-life data utilization and suggesting datasets for exploration. The survey facilitated building environment professionals’ voices, identified gaps between industry data and data science capabilities, and guided course improvements.

This insight extends beyond the course, aiding educators in tailoring content for professionals and potentially accelerating data science’s integration within the built environment context. This analysis informs discussion and suggestions on improving digital literacy and skill development in the building industry and the simulation community. The results showed that the gaps included the lack of data science application on real-life scenarios in various disciplines of the built environment – for example, application on BIM data in the design phase, project scheduling data in the construction phase, and HVAC data in the operations phase.

Acknowledgments

The authors would like to acknowledge the team involved in developing content for this course’s curriculum, including Ananya Joshi, Chun Fu, Mahmoud Abdelrahman, Matias Quintana, Miguel Martin, Sicheng Zhan, and Vanessa Neo. The course was supported by funding from the Learning Innovation Fund - Technology (LIFT) Grant at the National University of Singapore.

References


