Empowering Learners with a Low-Barrier Mobile Data Science Toolkit

Mobile Data Science Toolkit

Using MIT App Inventor to build data science mobile applications

Hanya Elhashemy[1,2], Robert Parks[1], David Y J Kim[1], Evan Patton[1], and Harold Abelson[1]

[1]Massachusetts Institute of Technology, [2]Technical University Of Munich, hanya.elhashemy@tum.de, rparks@mit.edu

This paper introduces a novel data science toolkit designed specifically for children, enabling them to create mobile apps integrated with data science capabilities. The toolkit showcases new features that simplify the data science process for young users. Additionally, the paper presents a collection of example apps created using the toolkit, highlighting the versatility and potential of this innovative platform. By empowering children to explore data science through app development, this toolkit opens exciting opportunities for hands-on learning and creative expression in the field of citizen science.

Keywords and Phrases: MIT App Inventor, Data Science, Mobile Applications, Citizen Science

1 INTRODUCTION

Many barriers, such as cost, access, and complexity, can prevent students from building working prototypes for performing authentic data science practices across the data science life cycle — collecting data, data cleanup, data visualization, and prediction (Mike, 2022). For example, online tools such as Python notebooks can be too complex for students new to Python, and such tools only allow users to perform parts of the data science life cycle. Moreover, few products enable students to build projects that visualize data on mobile platforms — which, after all, is where many scientists and citizen scientists collect and consume data visualizations today (Silvertown. 2009).

We discuss and present demos and student examples for a low-barrier mobile toolkit that allows learners to take part in the whole data science lifecycle. The demo and discussion are relevant for middle and high school teachers and facilitators of out-of-school programs. Participants can build simple mobile apps using MIT App Inventor and test these systems for themselves during the conference presentation.

The toolkit is a free suite of new, block-based-programming data science features that allow students to solve real-world problems that affect them or their community.

1.1 Motivation: Data Literacy

Data science is a multidisciplinary field that extracts insights from data by combining principles from mathematics, statistics, computer science, and domain expertise. Its applications span across various domains. We believe that introducing data science to children can boost their motivation for standard statistics and math school subjects. By showing them how these seemingly theoretical disciplines are applied in the real world to solve important problems, students can better grasp their significance. Data literacy is defined as the essential skills required to understand and use data effectively, including problem-solving, data collection, analysis, evaluation, visualization, and ethics. In today's digital age, data literacy is crucial for making informed decisions based on data.

In addition, data thinking involves a cognitive process that encourages individuals to approach problems and decision-making with a data-centric mindset, fostering skills like problem abstraction, pattern recognition, critical thinking, and

informed decision-making (Mike, 2022). The data lifecycle represents the entire journey of data, from creation to removal or archiving, and involves defining the problem, data collection, cleaning, exploration, machine learning, iterative evaluation, and publication of results. For learners, the data lifecycle serves as a helpful guideline, reducing the intimidation that comes with learning a new field like data science. Our goal with the toolkit features is to provide learners with a platform to go through the phases of the data lifecycle, acquiring essential 21st-century data thinking skills with ease and accessibility, while solving real-world problems of relevance to the learners.

1.2 IOT Connection and Data Management Features for Mobile Phone Apps

The App Inventor IoT component enables a connection between a mobile phone app and a low-cost Bluetooth sensor. This connection allows for real-time data collection from the sensor, ensuring accurate and reliable time series data acquisition. By leveraging Bluetooth technology, the app can connect to the sensor wirelessly, eliminating the need for complex wiring and enhancing the mobility of the data collection process (Lechelt, 2020; Clark, 2019). The online spreadsheet-to-mobile-device component streamlines data transport from spreadsheets to the mobile device. This functionality allows users to visualize, clean, and analyze the collected time series data conveniently. By integrating online spreadsheet services in the form of Google Sheets, the app facilitates easy access to data from anywhere, fostering collaborative data analysis and reducing the dependency on traditional desktop software. Additionally, cloud storage services can be integrated into the app, enabling seamless synchronization and backup of data to remote servers.

1.3 Anomaly Detection and Data Cleaning Techniques for Mobile Phone Data Analysis

The Anomaly Detection component allows for identifying anomalies and deciding whether to remove them based on the type of data and the user's domain knowledge. By incorporating advanced anomaly detection algorithms, the application can automatically flag potential anomalies in the data. However, it recognizes that not all anomalies are erroneous and may reveal valuable insights depending on the domain. Therefore, the feature allows users to exercise their domain knowledge and make informed decisions on whether to remove flagged anomalies or retain them for further analysis (Vaidya, 2023). The Preliminary Visualization feature enables users to perform initial visual analysis on their mobile phone's touchscreen. By providing interactive visualization tools, the application facilitates the identification and removal of out-of-context anomalies. Users can visually explore the data, observing patterns, trends, and irregularities directly on mobile devices. This feature empowers users to make data-driven decisions on removing visually distinguishable anomalies that do not align with the expected context of the data.

1.4 Mobile Charting for Visualizing Data in Mobile Applications

The Mobile Charting component provides young designers with options for visualizing data in a mobile application. By incorporating chart types such as bar graphs, line graphs, scatter plots, and pie charts, students can select the most suitable representation for their data based on the nature and characteristics of the dataset. This flexibility allows for effective communication and comprehension of the data, enabling students to identify patterns, trends, and relationships.

1.5 Linear Regression for Predictive Insights in Mobile Applications

The Linear Regression component allows students to add a line of best fit to showcase trends in the data. By fitting a straight line to the data points, they can observe the direction and magnitude of changes, helping them anticipate future outcomes. This feature helps students understand the relationship between variables and make predictions based on the linear trend. The feature also includes presenting statistical and probability information about the data. This information

aids users in assessing the strength and significance of the linear fit. For example, the feature provides correlation coefficients to quantify the degree of linear association between variables. These statistical insights enhance users' understanding of the data and assist them in making informed decisions.

2 EXAMPLES

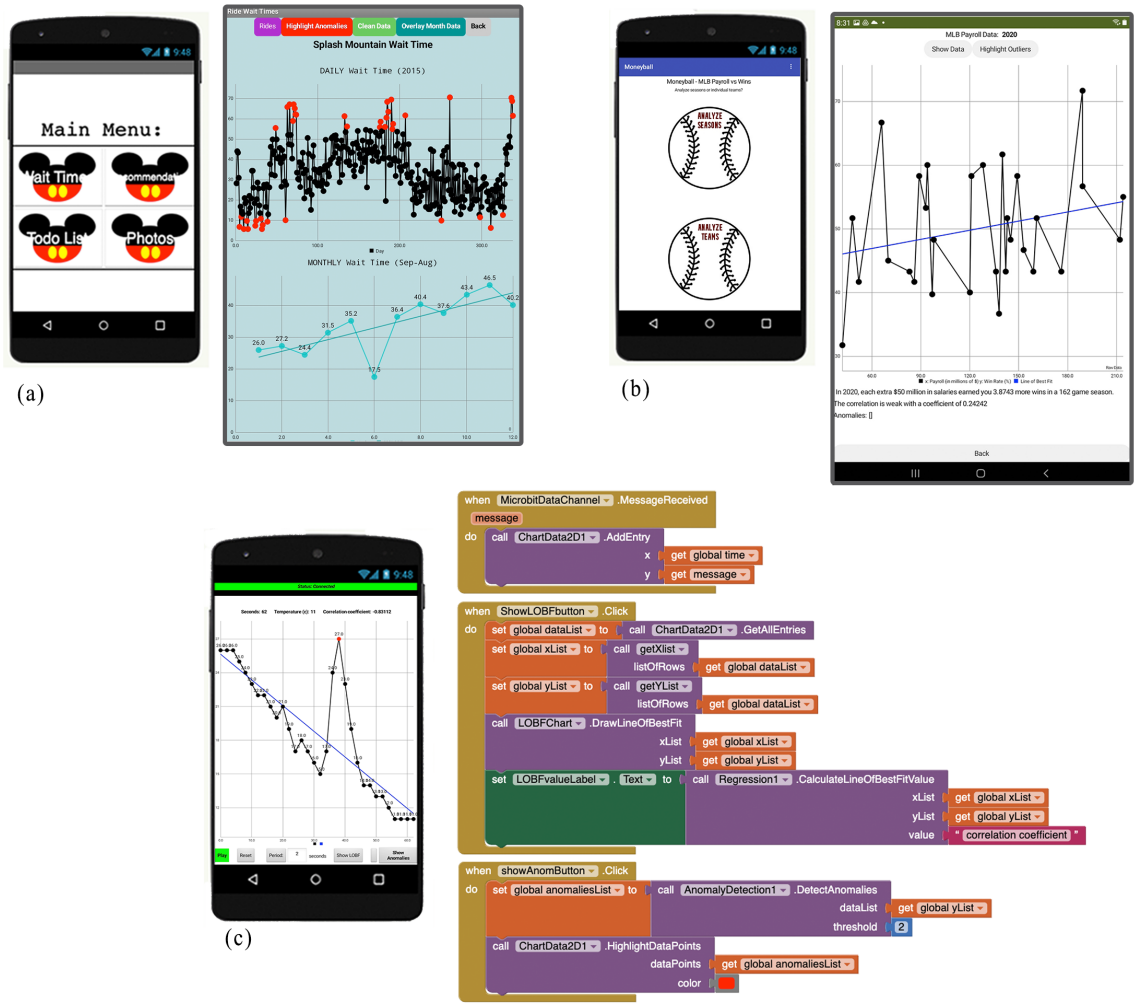

Figure 1: Images of example apps using the data science toolkit.

2.1 Analyzing Posted Wait Times for Disney World Rides: Insights for Average Wait Time Prediction

Figure 1 (a) shows an app for analyzing wait times for various rides at Disney World to determine the average time a child might have to wait in line. Raw data from a reliable source, specifically *touringplans.com*, provides information on ride wait times throughout the year. The data is graphed to identify trends in ride busyness during specific times of the day,

month, or year. Additionally, the app explores the significance of extremely low and high wait times, allowing children to understand factors such as weather conditions, park attendance, or ride maintenance that contribute to these outliers.

2.2 The Relationship between Payroll and Wins in Major League Baseball: An Anomaly Detection Approach

Figure 1 (b) investigates the relationship between team payroll and wins in Major League Baseball (MLB). Using data from over 150 games played by each team in a season, the app analyzes how spending on player salaries impacts a team's performance. It also acknowledges the presence of outliers, such as teams that overspend or underspend on player salaries, as exemplified in the movie *Moneyball*. Consequently, anomaly detection techniques are employed to identify and analyze such outliers, making this example suitable for anomaly detection in the context of MLB payroll and wins.

2.3 Gathering Time-Series Sensor Data: Temperature Readings from a Wireless IoT Sensor

Figure 1 (c) shows a mobile phone app with data gathered using a low-cost Bluetooth sensor. The data show a steady decrease in temperature that is briefly interrupted by a short-term increase (in this case, a warm coffee cup on the sensor for illustration). The line in blue is a line of best fit for the data, and the data point in red has been identified in the app as an anomaly based on a z-score greater than 2, as set in the App Inventor blocks, shown to the right.

App Inventor programming blocks show the App Inventor IoT component for receiving serial messages from the Bluetooth device, which are passed into the Mobile Charting component to generate the line graph. Blocks from the Linear Regression feature allow the user to draw a line of best fit in blue and calculate its correlation coefficient. Blocks from the Anomaly Detection component identify data points with a z-score greater than 2, then highlight them in red. By engaging children in hands-on activities and data tracking, this research aims to foster interest and understanding of the application of various sensors for directly gathering and analyzing data in many domains from sports performance to climate change.

3 CONCLUSION

By providing an accessible data science toolkit, this paper enables students to actively engage as data scientists in areas aligned with their interests and experiences. The inclusion of new features empowers young users to explore data-driven concepts, analyze real-world data, and create meaningful mobile apps that reflect their unique perspectives. This approach fosters a deeper understanding of data science and encourages students to apply their newfound skills and knowledge in diverse domains, promoting a sense of ownership and relevance in their data science journey.

ACKNOWLEDGMENTS

Jacky Chen, Tommy Heng, Gisella K Kakoti, Susan Lane, Evaldas Latoškinas, Ava G Muffoletto, Matthew R Quispe, Arianna C Scott, Isabela Naty Sanchez Taipe, and Jennet Zamanova

REFERENCES

Mike, K., Ragonis, N., Rosenberg-Kima, R. B., & Hazzan, O. (2022). Computational thinking in the era of data science. *Communications of the ACM*, 65(8), 33-35.

Silvertown, J. (2009). A new dawn for citizen science. *Trends in ecology & evolution*, 24(9), 467-471.

Lechelt, S., Rogers, Y., & Marquardt, N. (2020, June). Coming to your senses: promoting critical thinking about sensors through playful interaction in classrooms. In *Proceedings of the Interaction Design and Children Conference 2020* (pp. 11-22).

Clark, J., Falkner, W., Balaji Kuruvadi, S., Bruce, D., Zummo, W., & Yelamarthi, K. (2019, March). Development and Implementation of Real-Time Wireless Sensor Networks for Data Literacy Education. In *Proceedings of the 2019 ASEE North Central Section Conference, Morgan Town, West Virginia.* (pp. 22-23).

Vaidya, A., & Sharma, S. (2023). Anomaly detection in the course evaluation process: a learning analytics-based approach. *Interactive Technology and Smart Education.* Vol. ahead-of-print, No. ahead-of-print. https://doi.org/10.1108/ITSE-09-2022-0124