Puerto Rico Data Collection Web Platform: Puerto Rico Stats

José R. Martínez Torres
Master in Computer Sciences
Dr. Alfredo Cruz, Ph.D.
Electrical and Computer Engineering and Computer Science Department
Polytechnic University of Puerto Rico

Abstract *— Currently, the importance of data gathering and usage is at its highest point and constantly growing. Companies around the world are hiring data science teams to obtain and analyze massive amounts of data in order to provide the best services or products possible to their current or potential customers. Governments are also providing websites that conglomerate demographic/geographic data for public use. However, the Government of Puerto Rico falls a bit short on this due to the data often being extremely dirty, incorrect, and it being scattered across many websites that can be difficult to find. To ease some of the troubles of both searching for the data and knowing what to do with it, this project has provided a website that aggregates data from many sources related to Puerto Rico in one simple location. The project has also provided a section in which current or aspiring professionals can both obtain and learn how to utilize the data available, as well as providing resources for those that wish to learn more about the data science field.*

Key Terms — *data science, data accessibility, data readability, web platform*

INTRODUCTION

When making a decision, it is assumed that all available information has been taken into consideration in order to determine which decision would provide the most benefit or would be considered the "best" or "correct" option. With large-scale decisions particularly, all angles of information must be analyzed before coming to a final decision, starting with being able to access said information and all its variants. The same premise regarding data holds true for organizations and researchers, which rely greatly on accessing both public and private data in order to perform their work successfully. Statistics prove to be a tool used for calculated and informed decisions, its importance described as "so widespread, and the influence of statistics on our lives and habits so great, that the importance of statistics can hardly be overemphasized" [1].

Throughout this project, various topics, such as data accessibility and consumption and how these tie in together, will be explored in order to present the purpose of this project, detailing its completion process through phases. It will then dive into the problem of data accessibility and readability in the current world, where the internet has become commonplace. Finally, the project will present a web platform that aims to aid its users by providing data that is more accessible and readable for them to use to their advantage, help educate them on potential future decisions, or give them the necessary resources to learn more about the field of data science.

PROBLEM STATEMENT

The public availability of (mostly raw) data in Puerto Rico should, in theory, be a stepping stone for research led by corporations and individuals alike. All countries rely greatly on statistics for research development and public aid, requiring ample public data to be available. Currently, topics such as varied research studies are a widely used necessity for these countries' development [2, 3]. Hayslett suggests the government's use of statistics as well, for "the areas of taxation, funds spent for public works, public assistance funds, and so on" [1]. Puerto Rico widely aligns to this concept, providing a surplus of public data for anyone to use as they see fit. This benefits the general public by providing them with the necessary materials to potentially partake in important studies. However, there is one big problem with all this in the form of organizational lacking. Researchers must spend

much of their time digging through data made accessible in its existence, but inaccessible by the complications to actually get one's hands on it.

There doesn't seem to be a place, online or elsewhere, where anyone can go find out about all these sorts of resources that are available to the public with data visualization to accompany it. The sole exception is *Estadísticas.PR*, a site with some public polls and data that, while beneficial, is still lacking in volume of information and relevance. A common problem that plagues that platform is that, while it may have vast amounts of data from many sources, most of has not been updated in the last two years, while these datasets are expected to be updated monthly, causing users to leave the site in search for up-to-date data. This can be quite a problem because, while the website did provide the users with the initial knowledge of the dataset's existence, they stopped using the site and will most likely return to the platforms that provided them the updated datasets. All this information is lacking a means of accessibility, a more user-friendly hub for those who wish to delve into the data. Once users are able to access this data with ease, research conduction can skyrocket and open possibilities for more important developments to surge forward.

PROJECT GOALS

This projects' primary goal is to provide the general public, potential researchers, and aspiring/current data science professionals with a robust platform built as a web application in which they can view and download different statistics on Puerto Rico and also find resources on the data science field. The aim of this platform is to provide data accessibility to ease the strain on both researching corporations and individuals alike. To aid this mission, this platform will include pre-made examples with various forms of data visualization (such as charts, maps, or tables) for every entity offered on the platform, accompanied by links to the original data sources for further research. This hub aims to contain as much data on Puerto Rico as possible, from various sources located both within and outside of the island. Table 1 shows the currently available entities.

Table 1
Entities Available in the Platform

Entities Available on Website
Puerto Rico Electric Power Authority
Puerto Rico Economic Development Bank
Department of Education of Puerto Rico
Federal Housing Finance Agency
U.S. Bankruptcy Court
U.S. Bureau of Labor Statistics
U.S. Patent and Trademark Office
Puerto Rico Integrated Transportation Authority
Puerto Rico Trade and Export Company
Puerto Rico Department of Labor and Human Resources
Puerto Rico Institution of Statistics
U.S Social Security Administration
U.S. Census Bureau
National Oceanic and Atmospheric Administration

This will be achieved by offering links to the data's sources, as well as in-site explanations regarding the data's usability and applicability. Data will be presented in three forms:

- Raw unfiltered data, linked to its source. This will ensure that sources are properly credited and that original data methods are available to institutions or individuals attempting large-scale research or reference.
- Filtered, adjusted data. This can facilitate initial data searches and fit within the scope of an average person's research.
- Visual aids. As mentioned earlier, the use of charts, maps, tables, and graphics (accompanying filtered data) will be the final step in creating a highly accessible and user-friendly platform with public data.

In addition to all the data that is offered by the different entities on the platform, we have also included a blog section that contains resources valuable to those interested in the field of data science. These resources have a wide range of uses. The suggested platform would increase both the ease of access to the data provided by the

Government of Puerto Rico (and other entities) and the understanding of it with its various methods of visualization aiding the data, keeping user-friendliness and accessibility at the forefront of the project. This would make the platform a valuable tool for researchers to find the resources they need in order to conduct their studies.

Relevance and Significance

Companies have begun to rely more on data that is provided to them by users or other companies/sources, in an effort to provide a better product or better market themselves, for example. In the last decade alone, a 650% increase in data science roles has been observed [4]. With this, it is clear that there has been a massive boom when it comes to interest in the field of data science, suggesting that there will be data scientists willing to explore and learn about a much wider array of data. This is likely to promote new scientists to delve into their own country's data in an attempt to reveal information and improve quality of life and research standing.

Many areas of research have benefited from data sharing and accessibility. For instance, ecological data is seeing a rise in data sharing as more research sites and studies have begun to share resources or make their data available on websites for users to view and/or use in their own studies [5]. As the benefits of data sharing have become apparent in ecological academia, the belief follows with data science and the availability of general information.

Statistics have become increasingly more important than it already was in the data science field with the rise of prediction models [6], used in smartphones through features like Google Maps, Apple's Siri, or Amazon's Alexa; or, in the case of the real estate industry, for determining real estate prices, generating valuations and comparables, and for improvements in the field of affordable housing development [7]. Once again, these devices rely on both public and private accessible data for both their initial design and functionality updates, reaffirming the entire premise of data accessibility that this project is founded upon. Easy accessibility as a basic requirement for all public data is the first step towards enhancing and promoting scientific development across a myriad of research areas.

Methodology and Design

It is important to keep in mind that the readability and ease of use of the user interface is key to this platform's success. The users of the platform may range from data scientists with many years of experience to the general public with little to no research experience. Knowing this, the platform sticks to a simple design that is easy to follow while still maintaining an aesthetic that is easy on the eyes.

Content

Figure 1 shows a very simple diagram of the layout of the website's design. In the spirit of keeping the platform's design simple, the layout is separated into two sections: an always-visible vertical navigation bar with the platform's different sections (represented by the blue boxes with "Sec 1," etc.), and the contents of the chosen section to the right.

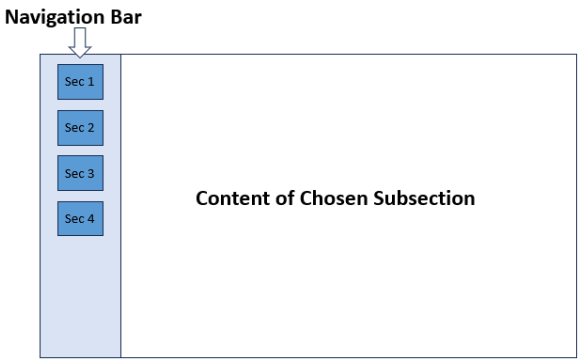

Figure 1
Simple Website Design

The first piece of content that the user is greeted with is the home section, which provides an explanation of the platform, its purpose and the content that can be found in it. The "main" content of the platform is the data related to Puerto Rico that is collected from the different government entities that are available to choose from, called the

Entity Data section. The second content is filled with resources for users interested in pursuing careers in data science, such as textbooks, certifications, job postings, and more. Every resource has links that the user can click to learn more. Table 2 shows the sections in which the content in the platform is separated into.

Table 2
Section Table

Section Number	Section Content
Section 1	Home
Section 2	Data
Section 3	About
Section 4	Blog

Database

In order to store all of the data that is being used in the platform, a database management system (DBMS) was necessary. The chosen DBMS is PostgreSQL, known for its scalability, stability, and reliability. We will also be using pgAdmin 4 as the platform to access the database that we will be using for the project. The Database is composed of five tables that contain all of the necessary information for the platform to run: Organization, OrganizationExamples, Datasets, BlogItems, and BlogItemTypes. An Entity Relationship Diagram (ERD) (figure 2) was made in order to better understand the relationship between the tables. Using the ERD, the database was created in pgAdmin running the necessary queries to create the tables and establish the relationships between the tables.

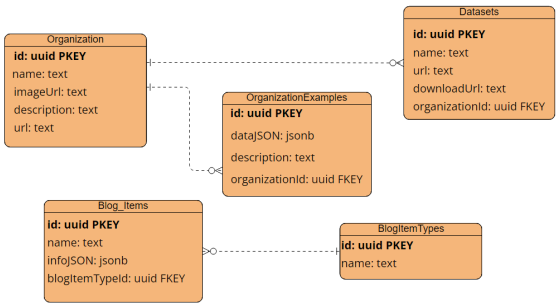

Figure 2
PRStats PostgreSQL Database ERD

RESULTS

This section will explain in detail the results of the completion of this project and how it relates to the initial idea that was proposed. The different aspects of the website will be shown below and explained in detail.

Home Section Results

The Home section is the landing page that the users see when they first enter the platform, which provides a simple explanation of the platform's purpose. Since dataset availability is the main purpose of the platform, there is a button on the middle-left section that will take you to the Entity Data section.

Entity Data Section Results

The Entity Data Section is the second and most important section of the platform. It begins with the users being provided a list of buttons with an image and name of each entity (figure 3). Clicking on one of the buttons takes the user to a page with the datasets available for download.

Entity Data Section: Entity Details subsection

As the Entity Data section is the most important part of the platform, readability and organization of this section was key. A diagram was made to lay out the structure that each entity would follow (figure 4). The Entity Detail subsection can be separated into three key sections: entity info, entity datasets, and entity examples. "Info" contains the entity's name, logo, a small description and a URL to access their website if there is one (figure 5). Entity datasets (figure 6) and examples (figure 7) are the most important sections. Here, as explained in the name, users can find the datasets available with the chosen entity, with a link to the page where it was obtained and a button to directly download the dataset. The examples provide a graph with some analysis on how it is read using the same data that is available for download from that entity.

Figure 3
Entities List Section of Platform

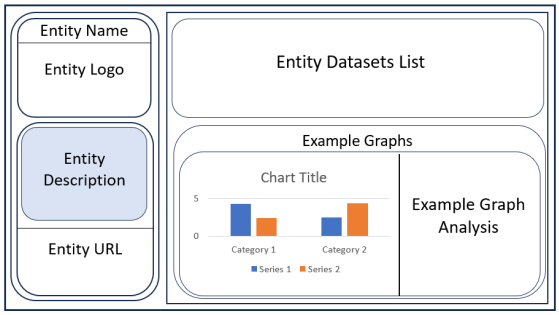

Figure 4
Diagram of "Entity Details" Subsection Layout

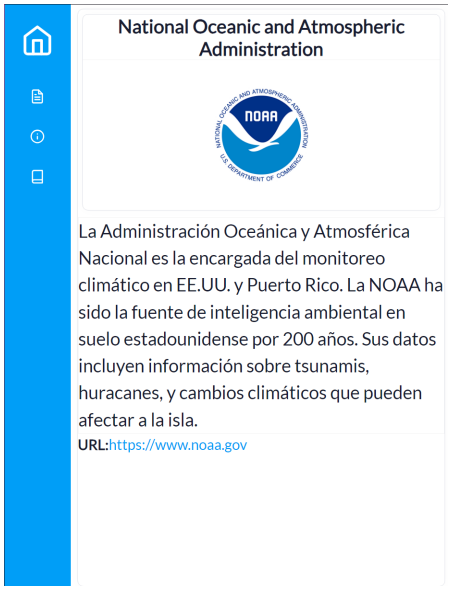

Figure 5
Entity Info sector of Entity Detail

Figure 6
Entity Datasets List

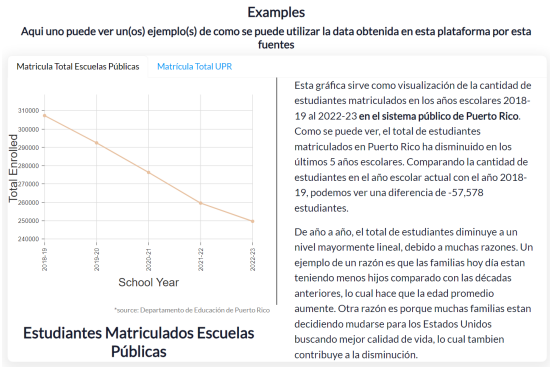

Figure 7
Examples section of Entity Details

Entity Data Section: Example Graphs

In order to highlight the utility of readily available data, included are some examples of graphs shown on the platform. These are made with the same data that is available to the users for download. These are accompanied with analysis and context to explain them, serving as a guide as to what can be done with and learn from the data.

Students Enrolled in School and University

The topic of education in Puerto Rico has been recurring during recent years in Puerto Rico. Whether it is the number of students enrolled, the lack of teachers, or the lack of support to repair schools or fulfill their operational needs, the educational system is a regular talking point of local news stations.

Over the last five years, Puerto Rico has seen a constant decline in the number of students enrolled at the beginning of the school year. Figure 8 shows the decline in K-12 public school enrollment in Puerto Rico. There is a myriad of reasons for this. The first and simplest reason is a decline in the population of Puerto Rico. The island's population has declined in the past 15 years as families have left the island in search for better opportunities. Because of this, most of the island's population is older, since they have already established their families on the island and cannot simply move because they wish to do so. Therefore, since most families leaving the island are the young ones, it is natural that the number of students enrolling in school is less. Another reason is there are less births daily compared to the 50s to 80s.

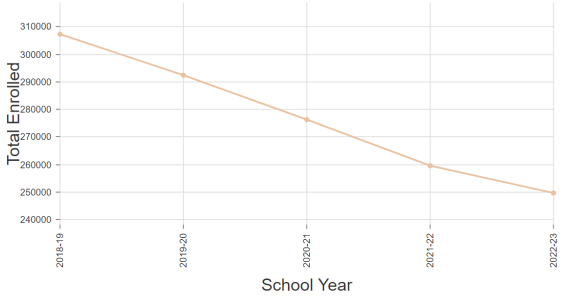

Figure 8
Public Schools Enrollment (K-12)

Besides enrollment for public schools, there is also student enrollment in the University of Puerto Rico, the island's largest university. Figure 9 shows enrollment in the university during the last five school years.

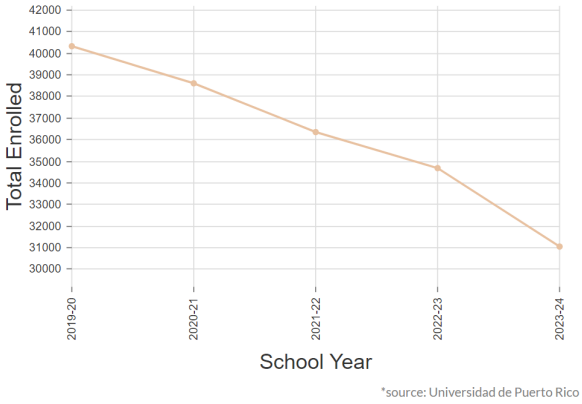

Figure 9
Graph of UPR Students per School Year

The numbers show that enrollment in public schools during the last five years is down 18.7%, and down 23% in the University of Puerto Rico. It is also projected that, in the next five years, the total could decrease up to 50% from the number of enrolled students in 2017 [8]. The University of Puerto Rico's sharper decrease is expected, as there are many other higher education institutions on the island. Young people are looking to study in the United States as some feel that their prospects will be higher outside of the island.

Sea Levels in Puerto Rico

Beaches are an incredibly important part of Puerto Rico and its culture. Many tourists travel every year to visit Puerto Rico and relax at the many beautiful beaches throughout the island. They are a great source of pride and have become a focus for both domestic and international tourism, which is one of the island's biggest economic sectors. Knowing this, protecting the environment and our beaches is critical. Therefore, relative sea level trends are a good point to study and analyze. Relative sea level represents the changes in the sea water levels relative to land with respect to the average conditions over a reference period. Figure 10 shows relative sea level trends in Puerto Rico.

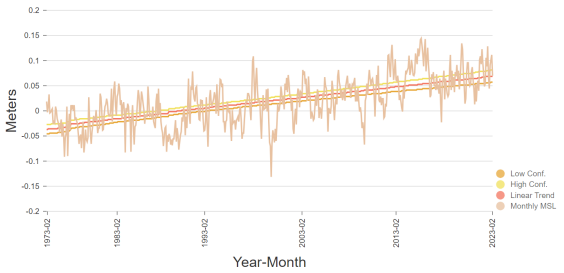

Figure 10

Relative Trend of Sea Levels in PR

The graph shows that, over the course of 50 years, relative sea levels have increased by around 0.1 meter, as per the linear trend. While this may not seem cause of concern in the immediate future, it can be quite worrying in the long term. Climate change has had a huge impact on the world, and the melting of the polar ice caps has increased the sea level around the world. Raising awareness of climate change is incredibly important in order to help combat it, in order to avoid the loss of our beaches. Due to this rise in sea levels, future storm surges, floods, and damages to coastal areas will be much more intense. Because of these damages in coastal areas, the population will be displaced and forced to seek safer homes [9].

Passengers in Urban Train System

Puerto Rico's metropolitan area's train line, Tren Urbano, or TU, is an urban railroad system inaugurated in 2004 that connects the populations from the municipalities of San Juan, Guaynabo, and Bayamón, with the idea of making public transportation more accessible in an island where having a car is considered a necessity. Ever since its inception, the TU is used by millions of passengers yearly. However, the amount of yearly passengers (figure 11) is not enough to outweigh the perceived negatives and the cost required to maintain it. The project's construction was marked by multiple delays and budgetary deficits that resulted in the system costing much more than was projected for a smaller system than originally planned due to budgetary constraints.

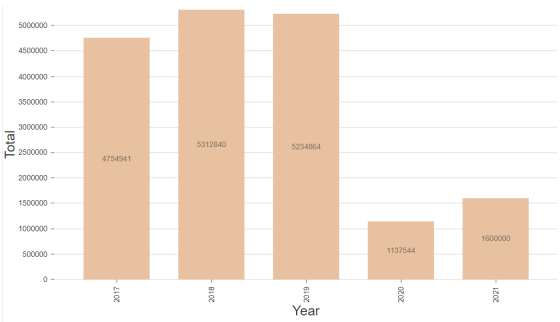

Figure 11

TU Yearly Passengers

Due to the system being a constant talking point in the island for the last 25 to 30 years, its inclusion in the platform was key. During the last five years, the amount of TU yearly passengers has declined like never before. In 2017 to 2019, usage saw an uptick, with an average 5.1 million yearly passengers. Problems arose in 2020, when usage took a nosedive of 4 million fewer passengers in a year. Notably, the TU was closed most of the year due to the COVID-19 pandemic. Nevertheless, the following, 2021, post-pandemic numbers did not increase much. In addition, while not recorded on figure 11, previous years also saw a downward trend; yearly passengers in 2014 were 8.4 million [10].

The pandemic is not the main culprit of the TU's decline. However, it seems to have been the straw that broke the camel's back. Added to its poor planning, the absence of connections between the train and public bus systems, and the overall abandonment of spaces in the facilities that were expected to be private establishments have caused the TU to never manage to reach the projected income that was necessary to justify its cost.

About Section Results

This section details the purpose of the platform and why it was created, which is to provide its users with clean and accessible data in order to facilitate the access and understanding of the information provided by the entities supported on the platform.

Blog Section Results

When implementing the blog section to the platform, the answer to what resources should be provided changed many times. Since the main

purpose of the blog is to provide current/aspiring professionals with some resources that might aid them in learning about different aspects of the field, it would also be wise for the different resources to be tailored for different professional sectors. The sections have been separated into two subsections (figure 12) in order to provide information as efficiently as possible, improving readability.

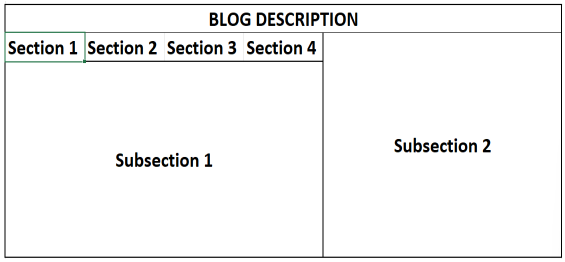

Figure 12
Diagram of Blog Section Layout

The blog section is divided into four subsections of resources that could be useful for data science professionals (table 3). The reason for the different types of resources is to provide broader utility to more users instead of being more focused on a single resource that could have a more niche interest rate.

Table 3
Subsections Table of Blog

Subsection Number	Subsection Content
Subsection 1	Research Papers & Books
Subsection 2	Blogs & Articles
Subsection 3	Job Postings & Certifications
Subsection 4	Educational Institutions

The purpose of this is to provide more ways in which the platform can be useful to its users. Even if the users end up not using any of the data provided, they can still find something useful out of the multiple resources that can be found in the blog section.

Research Papers and Books Subsection

These are links to research papers that have been published in the past year (since 2022) and in-depth and informative books pertaining to different aspects of the data science field—such as big data, deep learning, and data mining—intended for those who are interested in reading more in-depth published work or are searching for references for their own studies/papers. Figure 13 shows the subsection's structure. Examples of how the research papers and books are displayed in the subsection can be seen in figures 14 and 15.

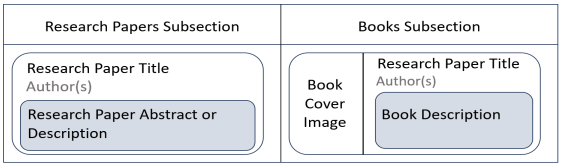

Figure 13
Diagram of Research Papers and Books subsection

Figure 14
Example of Research Paper in Subsection

Figure 15
Example of Book in Subsection

Blogs and Articles Subsection

The Blogs and Articles subsection is for those who like to casually read articles on computer science or data science. Other blogs in the field are recommended, such as *Data Science 101*, a blog that helps its readers hone their data science skills with advice, resources, and news. This section also provides potentially interesting links to some articles from the previous year from different sources. Figures 16, 17, and 18 show the structure of this subsection.

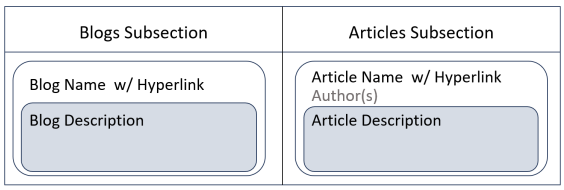

Figure 16
Diagram of Blogs and Articles Subsection

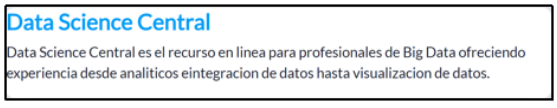

Figure 17
Example of Blog Entry Subsection

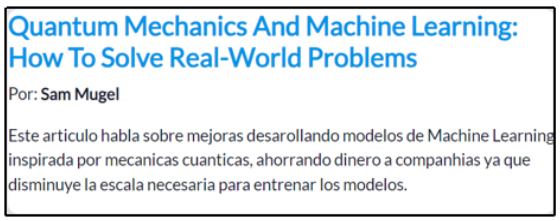

Figure 18
Example of Article Entry in Subsection

Job Postings and Certifications Subsection

For professionals who are looking for resources that are not just reading, there is a section that offers links to data science certifications that would enhance any resumé, and links to current data science job listings in Puerto Rico. Below are examples of this section's structure (figure 19) and examples of job postings (figure 20) and certifications (figure 21).

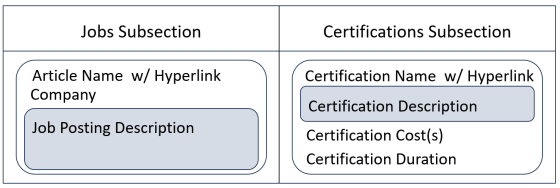

Figure 19
Diagram of Job Postings and Certifications Subsection

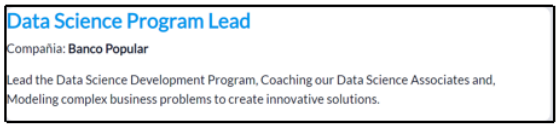

Figure 20
Example of Job Posting Entry in Subsection

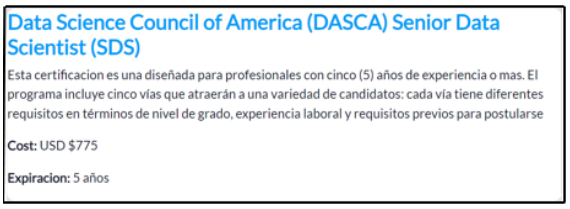

Figure 21
Example of Certification Entry in Subsection

Educational Institutions Subsection

The Educational Institutions subsection is made for undergraduate and postgraduate students and those who are thinking of pursuing data science careers; it offers information on higher-education institutions offering under- or post-grad tracks for data science. Figure 22 shows how the subsection is structured, with the list of bachelor's degrees on top. Each item contains the name of the degree hyperlinked to their universities' web pages and a short description. In addition, there is a button next to the title for downloading any degree's curriculum. Figures 23 and 24 show examples of items from institutions on Puerto Rico, as seen on the platform.

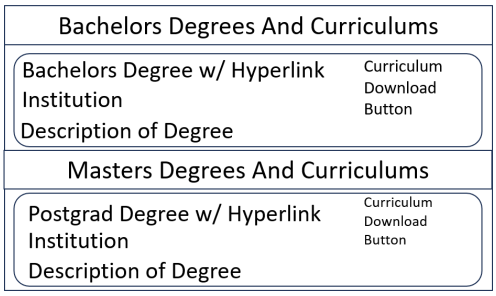

Figure 22
Diagram of Educational Institutions Subsection

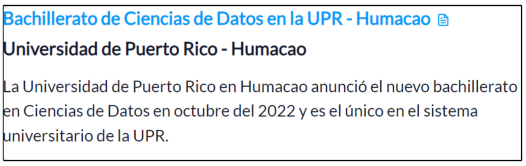

Figure 23
Example of Bachelor's Degree Entry in Subsection

Figure 24
Example of Master's Degree Entry in Subsection

Future Work

The project's initial objective has been met with the completion of the platform. For future work, more entities could be added over time and existing ones could be expanded. The ability to view the datasets on the platform with native table support is another feature that could be added in the future. A comments field in the blog section could be added for the platform's users to interact with the article and each other as the platform begins to expand and publish its own data science articles. Another way to bring more users could be to make the website's

layout and presentation more modern. Maintaining a sleek, professional design adds more credibility to the platform, which may entice new users to be return-users.

CONCLUSION

The purpose of carrying out this Puerto Rico Stats Platform project is to contribute to the field of data science in Puerto Rico by providing current professionals with datasets related to the island and aspiring professionals with resources that they can utilize to learn more about the field if they plan on pursuing a career in it in the future. With this in mind, research was done in order to determine how the platform would be laid out and what information would be provided in order to achieve a balance of information that is useful to both current and aspiring professionals. For the data that is being provided, research was done to determine what a respectable entity would include in the platform, to ensure that the data provided could be easy to obtain, understand, and implement into studies.

The field of data science is constantly growing, and demand for it is increasing every day as more and more companies are creating data science teams, so the demand for data scientists is constantly on the rise. Therefore, the more information we can provide to interested individuals in order to help them on their path to becoming data scientists, the better. However, a balance must be struck on what resources are provided to help these aspiring professionals learn, as these may be in different stages of their career or may choose to learn with different resources. Thus, the balance was struck between providing textbooks and research papers for the more academic types, and blogs and articles for those who are looking for a more “casual” approach. Whether a user is a college student looking to dip their toes into the field of data science or a long-time professional, there is something for everybody in the Puerto Rico Stats Platform.

REFERENCES

[1] H. T. Hayslett and P. Murphy, *Statistics*. London: Made Simple Books.

[2] S. I. Doguwa, "Statistics for national development," *CBN Journal of Applied Statistics*, vol. 1, no. 1, pp. 99–106, 2010. Available: http://hdl.handle.net/10419/142039

[3] PARIS21, "Counting down poverty: The role of statistics in world development," Accessed: August 29, 2022. Available: https://paris21.org/sites/default/files/2532.pdf

[4] New Jersey Institute of Technology, "The data boom spurs new demand for data scientists." Accessed: August 30, 2023. Available: https://sponsored.chronicle.com/the-data-boom-spurs-new-demand-for-data-scientists/index.html

[5] D. P. C. Peters *et al.*, "Long-term trends in ecological systems: A basis for understanding responses to global change," US Dept. of Agriculture, September 2013. Available: https://www.researchgate.net/publication/260248230_Long-Term_Trends_in_Ecological_Systems_A_Basis_for_Understanding_Responses_to_Global_Change

[6] V. Jalajakshi and A. n. Myna, "Importance of statistics to data science," in Global Transitions Proceedings, International Conference of Intelligent Engineering Approach, Raichur, Karnataka, India, 2022. Available: https://doi.org/10.1016/j.gltp.2022.03.019

[7] S. J. Park, "Data science strategies for real estate development," M. S. thesis, Program in Real Estate Development, MIT, Massachusetts, 2020. Available: https://hdl.handle.net/1721.1/129099

[8] L. Pineda Dattari, "Descenso de matrículas en universidades implicaría más del 50% de lo que había en 2017," *Noticel*, April 13, 2023. Available: https://www.noticel.com/educacion/ahora/top-stories/20230413/descenso-de-matriculas-en-universidades-implicaria-mas-del-50-de-lo-que-habia-en-2017/

[9] National Geographic Society, "Sea level rise," May 20, 2022. Available: https://education.nationalgeographic.org/resource/sea-level-rise/

[10] J. O. Delgado Rivera, "El Tren Urbano pierde 4.8 millones de pasajeros en cinco años," *El Nuevo Día*, May 18, 2019. Available: https://www.elnuevodia.com/noticias/locales/notas/el-tren-urbano-pierde-48-millones-de-pasajeros-en-cinco-anos/