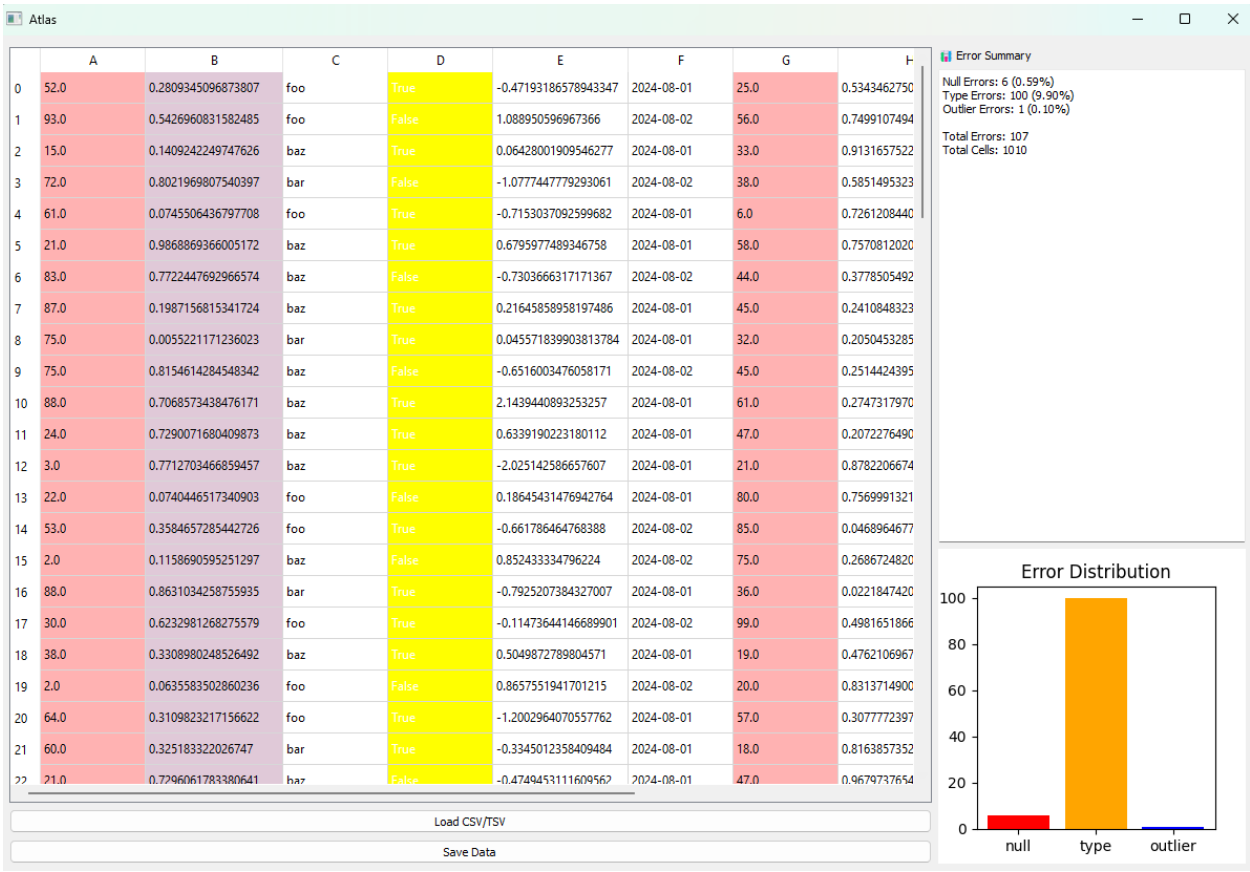
Atlas



	A	B	C	D	E	F	G	H
0	52.0	0.2809345096873807	foo	True	-0.47193186578943347	2024-08-01	25.0	0.5343462750
1	93.0	0.5426960831582485	foo	False	1.088950596967366	2024-08-02	56.0	0.7499107494
2	15.0	0.1409242249747626	baz	True	0.06428001909546277	2024-08-01	33.0	0.9131657522
3	72.0	0.8021969807540397	bar	False	-1.0777447779293061	2024-08-02	38.0	0.5851495323
4	61.0	0.0745506436797708	foo	True	-0.7153037092599682	2024-08-01	6.0	0.7261208440
5	21.0	0.9868869366005172	baz	True	0.6795977489346758	2024-08-01	58.0	0.7570812020
6	83.0	0.7722447692966574	baz	False	-0.7303666317171367	2024-08-02	44.0	0.3778505492
7	87.0	0.1987156815341724	baz	True	0.21645858958197486	2024-08-01	45.0	0.2410848323
8	75.0	0.0055221171236023	bar	True	0.045571839903813784	2024-08-01	32.0	0.2050453285
9	75.0	0.8154614284548342	baz	False	-0.6516003476058171	2024-08-02	45.0	0.2514424395
10	88.0	0.7068573438476171	baz	True	2.1439440893253257	2024-08-01	61.0	0.2747317970
11	24.0	0.7290071680409873	baz	True	0.6339190223180112	2024-08-01	47.0	0.2072276490
12	3.0	0.7712703466859457	baz	True	-2.025142586657607	2024-08-01	21.0	0.8782206674
13	22.0	0.0740446517340903	foo	False	0.18645431476942764	2024-08-01	80.0	0.7569991321
14	53.0	0.3584657285442726	foo	True	-0.661786464768388	2024-08-02	85.0	0.0468964677
15	2.0	0.1158690595251297	baz	False	0.852433334796224	2024-08-02	75.0	0.2686724820
16	88.0	0.8631034258755935	bar	True	-0.7925207384327007	2024-08-01	36.0	0.0221847420
17	30.0	0.6232981268275579	foo	True	-0.11473644146689901	2024-08-02	99.0	0.4981651866
18	38.0	0.3308980248526492	baz	True	0.5049872789804571	2024-08-01	19.0	0.4762106967
19	2.0	0.0635583502860236	foo	False	0.8657551941701215	2024-08-02	20.0	0.8313714900
20	64.0	0.3109823217156622	foo	True	-1.2002964070557762	2024-08-01	57.0	0.3077772397
21	60.0	0.325183322026747	bar	True	-0.3345012358409484	2024-08-01	18.0	0.8163857352
22	21.0	0.7296061783380641	baz	False	-0.4749453111609562	2024-08-01	47.0	0.9679737654

Team Members:

Avery Willets, Data Science

Technical Field

This project involves the technical fields of data collection, data preparation, and data visualization.

Background Information

This project was inspired by a problem I consistently run into when performing analysis: Preparing the data takes too long, and I often struggle to quickly grasp what is wrong within a dataset. Thus, I decided to create a tool to help people like me explore data!

*Prior Art (legal term)

Alteryx: Alteryx is a data preparation and analysis tool. It offers several services for data preparation, such as input tools to connect to data sources and cleaning tools to filter and sort data. Atlas differentiates itself from Alteryx by focusing on highlighting errors specifically and by implementing a unique flag system and filter function design.

URL: https://www.alteryx.com/platform-services

Tableau Prep: Tableau Prep is a data preparation tool from Tableau (an industry-standard data science service) that focuses on data preparation and cleaning within Tableau applications. Atlas differentiates itself from Tableau Prep by existing as its own standalone program. This may be subject to change, as Atlas may become an extension for another data science tool. Atlas also differentiates itself from Tableau Prep by providing users with a unique flag system that highlights errors.

URL: https://help.tableau.com/current/prep/en-us/prep_about.htm

Microsoft Power BI Power Query: Microsoft Power BI (another industry standard data science service) contains data preparation capabilities through its Power Query feature. Power Query can take data from various data sources and be used to clean it. Atlas differentiates itself from Power Query by not being connected to the Microsoft umbrella nor the Power BI system, and by providing users with a unique flag system that highlights errors.

URL: https://learn.microsoft.com/en-us/power-query/power-query-ui

Project Description

Atlas accelerates the process of collecting and preparing data by visually flagging errors within a dataset. It contains flexible filters to show a variety of those errors depending on user choice. Atlas aims to empower data analysts through this efficiency to reduce time spent cleaning and preparing data and more time uncovering insights from that data to inform business decisions.

Innovation Claim

Atlas is innovative because it accelerates the process of preparing data more by providing users with visual shortcuts in the form of flags around errors within their datasets.

Usage Scenario

One usage scenario of the Atlas framework is to visually organize data according to a specific feature within the dataset. For example, when analyzing a dataset about internet traffic in a cyber security setting, a user could map the data from different IP addresses to their own respective colors. Then, when evaluating the data based on another feature such as the port used or service used, the user can still quickly and easily identify those IP addresses based on their colors.

Evaluation Criteria

The following questions will identify the successful completion of this project.

Can the user interact with Atlas's interface? (Y/N)

Can the user filter data according to errors? (Y/N)

Can Atlas present the filtered data with visual flags around the proper errors? (Y/N)

Can the user successfully upload their file to Atlas? (Y/N)

Can the user successfully download their file from Atlas? (Y/N)

Objectives and Tasks Associated with the Project

Objective 1: Atlas has an interactable user interface.

Task 1.1: Build a grid-based GUI that the data will be put into. (Python, PyQt)

Task 1.2: Implement an example or dummy dataset into the program and ensure the system can present it. (Python, pandas, NumPy)

Task 1.3: Implement a blank data filtering function template and button to run it on click and ensure they work. (Python, pandas)

Objective 2: Atlas can visually highlight errors for users.

Task 2.1: Use the template to create a function for NaNs and decide on a color for it.

Task 2.2: Use the template to create a function for values of wrong datatype and decide on a color for it.

Task 2.3: Use the template to create a function for duplicate values and decide on a color for it.

Task 2.4: Use the template to create a function for values outside 3 standard deviations from their respective columns' mean and decide on a color for it.

Task 2.5: Use the template to create a function for wrong column length and decide on a color for it.

Task 2.6: Map the functions to the output GUI and ensure correctly Atlas presents errors visually.

Objective 3: Atlas can read from and write to user .csv files.

Task 3.1: Implement a file-reading function to interpret uploaded .csv files.

Task 3.2: Implement a file-uploading function for the user.

Task 3.3: Attach the file-uploading function to a button the user can interact with.

Task 3.4: Implement user input to allow them to modify the document.

Task 3.5: Implement error handling in user modification functions.

Description of Design Prototype

Atlas operates using a GUI (Graphical User Interface). It is bult with the language Python, with libraries including Pandas, NumPy and PyQt. It is built primarily within VSCode. Google Colab was also used in several phases of creation.

The prototype is a small GUI that requires a .csv file to be fed into it before operating. The user can select which filters they'd like to implement from among 3 selections, and then the prototype presents the data with visual flags outlining the filtered data. This is intended to demonstrate the ability to read a file, interpret its contents, and accelerate data preparation by clearly communicating insights to the user.

Evaluation Plan

Atlas will first be evaluated on its user interface. This will be done by booting up Atlas to demonstrate the graphical user interface. Then the user will press a button to show that the interface can be used. Atlas will then be evaluated on its ability to visually highlight errors for users. This will be done through demonstrating Atlas's visuals over sample datasets with different errors. Finally, Atlas will be evaluated on its ability to read and write to files. This will be assessed through a concluding demonstration in which Atlas is started and tested with a new .csv file, determining whether Atlas can parse the data from the file.

Note: This section must be revised prior to PRO483 to describe the full evaluation plan as it was actually implemented.

Project Completion Assessment

Note: This section must completed prior to PRO483.

Provide an in-depth description of the completion assessment of your project. Describe how well the completed components function and highlight the innovative facets of your design. This is sometimes known as a "Post-Mortem" or "Lessons-Learned Report". A good approach for this section is to answer the following 4 questions: "What went right? What went wrong? What was learned throughout the process? What would be done differently if you had to do it again?

Appendices

Appendix A: Prototype Program File – AtlasSystem.py

Appendix B: GUI Framework File – GuiFramework.py

Appendix C: Error filter functions file – FilterFunctions.py

Appendix D: Dummy dataset – dummy_dataset.csv