April 10, 2023

**C964: Computer Science Capstone**

By: Austin Kim

Task 2 – Parts A, B, C, D, E

[Part A: Project Proposal for Business Executives 3](#_Toc132026372)

[Letter of Transmittal 3](#_Toc132026373)

[Project Recommendation 4](#_Toc132026374)

[Problem Summary 4](#_Toc132026375)

[Application Benefits 4](#_Toc132026376)

[Application Description 5](#_Toc132026377)

[Data Description 5](#_Toc132026378)

[Objectives and Hypothesis 6](#_Toc132026379)

[Methodology 7](#_Toc132026380)

[Funding Requirements 8](#_Toc132026381)

[Data Precautions 8](#_Toc132026382)

[Developer’s Expertise 8](#_Toc132026383)

[Part B: Project Proposal 9](#_Toc132026384)

[Problem Statement 9](#_Toc132026385)

[Customer Summary 9](#_Toc132026386)

[Existing System Analysis 9](#_Toc132026387)

[Data 10](#_Toc132026388)

[Project Methodology 10](#_Toc132026389)

[Project Outcomes 11](#_Toc132026390)

[Implementation Plan 11](#_Toc132026391)

[Evaluation Plan 12](#_Toc132026392)

[Resources and Costs 12](#_Toc132026393)

[Timeline and Milestones 13](#_Toc132026394)

[Part C: Application 14](#_Toc132026395)

[Part D: Post-implementation Report 14](#_Toc132026396)

[A Business (or Organization) Vision 14](#_Toc132026397)

[Datasets 15](#_Toc132026398)

[Data Product Code 16](#_Toc132026399)

[Objective (or Hypothesis) Verification 18](#_Toc132026400)

[Effective Visualization and Reporting 18](#_Toc132026401)

[Accuracy Analysis 20](#_Toc132026402)

[Application Testing 21](#_Toc132026403)

[Application Files 22](#_Toc132026404)

[User Guide 22](#_Toc132026405)

[Summation of Learning Experience 27](#_Toc132026406)

[Part E: Sources 29](#_Toc132026407)

# Part A: Project Proposal for Business Executives

## Letter of Transmittal

April 10, 2023

Gigi Ruten, CTO

TastyFish

111 Freedom Road

Dallas, Texas, 92922

Dear Mrs. Ruten,

Our organization faces a critical problem. The time it takes for our fish to travel from fishing boats to clients’ warehouses is too long. We all know that due to the unstable motion of fishing boats, fishers cannot weigh fish individually. The sea’s waves prevent accurate weight measurements. Our problem lies in the company’s method of weighing fish. Currently, it is letting each fish dry and then placing each fish on a scale. This proves to be inefficient because the shelf life of the fish lowers when waiting for it to dry and also during the weighing process.

A solution we would like to recommend is to implement a prediction system that learns from previous fish data and calculates the weight of all our future catches. Since our fishers are already required to measure the five length dimensions of each fish they catch, we could use that information to predict its weight. This is possible using a method in computer science called “machine learning.” Machine learning is when a computer can mimic human intelligence and learn from a set of data. The computer is then able to solve a problem without being told exactly what it has to do.

This solution would benefit TastyFish because it would completely remove the need to dry and weigh each fish individually. Machine learning would be able to predict the weight of the fish using only the five length measurements. This would improve our throughput, labor needs, time, and costs. Since selling fish is what we do, this solution would essentially let us bring the fish to clients in less time.

Our team has calculated an estimated total cost to fully implement this solution into our company. The cost would be $536, 350. This cost includes all hardware, software, data, staff, and upkeep used during the creation and maintenance of the solution.

Having studied machine learning and software development in my undergraduate studies, I have developed a strong ability to complete this task. My experience both inside and outside the classroom has prepared me for this project. My 2.5 years of internship experience have developed my ability to work in a professional environment. The completion of my bachelor's degree will be soon and having completed most of my coursework, I believe I had sufficient knowledge to accomplish this task.

I hope that our solution will be turned into a real product that saves our company’s stagnation. Thank you for your time in advance. I hope to hear back from you soon.

Sincerely,

Austin Kim



## Project Recommendation

### **Problem Summary**

The Project:

* The project will include gathering currently existing data about fish and converting that into something that our machine learning tool can learn from. We will use the weight of the fish and also five different length measurements in the training process. Our machine learning tool will use math to teach itself how to predict the weights. Once we have trained our machine learning tool to accurately predict the weight of a fish using that data, we will create a user-friendly interface that can communicate with the trained tool. The interface will allow users to use the tool and input length measurements to receive a predicted weight.

Setting & Need for Project:

* The current setting that TastyFish is in shows a huge need for this project. TastyFish is in a circumstance where it is bleeding money during the fish processing stage. Other competitors have adopted advanced technologies and robotics to automate and enhance their businesses. On the other hand, TastyFish remains in a state of stagnation. TastyFish will soon fail if it does not dramatically evolve its business processes to keep up with the competition.
* This project is needed because TastyFish is risking the company’s profits and prosperity through its inaction. The solution would provide an immense increase in the supply, quality, and delivery of fish. This boost in TastyFish’s core product would provide TastyFish with enough momentum to bring large profits for many years.

Business Needs:

* This project addresses TastyFish’s business needs because TastyFish’s core mission is to provide quality fish to its clients. Factors that contribute to the success of this mission include the rate of fish being delivered and also the freshness of the fish. Both aspects are addressed in the proposed solution because the machine learning tool allows the weight of the fish to be predicted without the fish sitting in nonideal conditions.

Delivery Goals:

* This project aims to deliver a tool that predicts the weight of the fish using machine learning. This solution will achieve results that allow TastyFish to improve its business operations and lead to an overall boost in productivity, quality, and profits.

### **Application Benefits**

Business Needs:

* As mentioned before, “This project addresses TastyFish’s business needs because TastyFish’s core mission is to provide quality fish to its clients. Factors that contribute to the success of this mission include the rate of fish being delivered and also the freshness of the fish. Both aspects are addressed in the proposed solution because the machine learning tool allows the weight of the fish to be predicted without the fish sitting in nonideal conditions.”
* In addition, TastyFish is required to weigh the fish for tracking its inventory. It is crucial to track certain data about fish as it can help lead to discovering certain trends. An example of a trend would show fish are heavier during summer seasons due to the increased food supply. Also, the weight of the fish is critical when considering the client's need for such details. Our clients are mass buyers of fish and so our business operations require us to weigh much fish at all times of the year. This solution will meet the needs of the business because it will facilitate the integrity of records and inventory information. The trends that are studied will be more readily available as it becomes easier to weigh much fish at once.

Benefit from the Solution:

* The business will benefit from the implementation of the machine learning tool because of the way it allows many other processes. Instead of wasting time, energy, and resources on weighing the fish, the preparation and delivery of the fish can occur early. This will allow TastyFish to sell a fresher product that arrives to clients directly from the sea and in a short amount of time.

### **Application Description**

Technical Details:

* The application will solve the problem by using machine learning. The problem is that the weight of fish needs to be predicted using the five length measurements that exist. The application will contain a method of predicting the weight of a given fish using machine learning.
* The machine learning tool will be an offline application that can be utilized through any text interface. When the user starts the application, the user will navigate the menu to reach the input section. Here, they will enter the five measurements from any fish into the application. The application will then return the predicted weight.
* The technical details are mostly in the internal steps that happened after the input of lengths and before the output of the weight. In these steps, the tool that can predict the weight exists. The tool will be made to learn from existing fish data. The exact way it does this is by using complex mathematical calculations to predict the weight of a fish using various real measurements. The formula for calculating the prediction changes itself with each success and error. The tool does this repeatedly until it can achieve a prediction accuracy that is suitable for this task. This is how the tool will be created.
* Now, the inputted measurements are given to the tool. The tool takes these numbers and uses its customized formula to calculate a predicted weight. The weight is then given to the user.
* These steps outline how the application will solve the problem. The problem is that the fish need to be weighed more efficiently. The application is a tool that can calculate the weight of a fish without having to wait for it to dry or place it on a scale. Thus, the technical details of this application prove that the tool can accurately predict the weight of the fish and solve the problem of needing to manually weigh each fish on a scale.

### **Data Description**

Origins of Raw Data:

* The raw data will be sourced from a data set from this link (<https://www.kaggle.com/datasets/aungpyaeap/fish-market>).
* The raw data consists of the various fish. The data includes the fish species, weight, and five length measurements.

Type of Data:

* The weight and the five length measurements are quantitative data. Each of these six data points is a number represented by centimeters or grams.
* The species of the fish is considered nominal data. This is because the species of a fish does not correlate with any number and does not correlate with any certain order.
* The data structure which holds all this data can best be described as a 2-D array. An array is an ordered list of objects. A 2-D array is an ordered list of arrays. This would best describe the data.
* However, the raw data is stored in a CSV file which means that each value is separated by commas and each set of values is separated by rows.
* This format follows the conventional format that Excel files currently use to store information.

Variables:

* The independent variables include the 5 length measurements and the species name.
* The dependent variable is the weight of the fish.

Anomalies & Limitations:

* Anomalies in the data include one of the fish, which is a Roach, weighing 0 grams but having length measurements of 19, 20.5, 22.8, 6.4, and 3.3. Other anomalies in the data include examples where a specific numerical data point is strangely low or high compared to the other values.
* Limitations in this data include the lack of entries. With less than 200 sets of fish measurements, it will be somewhat less accurate due to the lack of training data.

### **Objectives and Hypothesis**

Desired Outcomes:

* The desired outcome of the project is to have a functioning application that uses machine learning to predict the weight of a fish given five length measurements. To explain, the application will hopefully contain a machine learning tool that will be trained using sufficient data. The data will hopefully be credible data containing previously inputted fish weights and length dimensions. The tool will hopefully be able to output the predicted weight accurately and quickly.

Hypothesis:

* The hypothesis of the project is “The application will use a trained machine learning tool to accurately predict the weight of a fish when given five length measurements.”

Desired Prediction Accuracy:

* In all cases, the prediction accuracy should be as close to 100%. It is sometimes impossible to achieve this but the goal for this project is to achieve a prediction accuracy rating of 80%.

### **Methodology**

Development Methodology:

* The methodology used to develop and implement this project will be the Agile methodology.
* The Agile methodology is appropriate because our solution will involve the creation of a user interface, the creation of a machine learning tool, the training of the machine learning tool, and the creation of software to support the interface. These steps do not need to be done in sequential order. The training of the machine learning tool can start simultaneously with the designing of the user interface. Because the formula that the machine learning tool uses will need to be tweaked, it is important that our project development methodology uses an iterative approach so that the machine learning tool can constantly be improved. As the tool improves, the team must have the ability to use feedback from TastyFish to implement any changes. The core significance of the Agile approach is that it allows the project to receive constant feedback from the stakeholders and constantly improve the product in each step. This is important because there will be the development of a user interface, which will be catered to how TastyFish would like it to look. The constant improvement of the application will help create a stronger and more accurate prediction tool.

Agile Methodology Phases:

* To explore the phases of the Agile Methodology, we will use an article, titled “Beginner’s Guide to Agile Project Management", which has been published by Adobe (Adobe Communications Team).
* First is the “Project Planning" phase. Here the project's overall purpose, goal, value, etc. will be determined. The scope of the project is determined but because it is Agile, all these things can be changed if necessary in the future.
* Second is the “Product Roadmap Creation” phase. Here, the project backlog is created that has all the deliverables that will be created in each sprint.
* Third is the “Release Planning” phase. Due to Agile’s iterative nature, the sprints will be releasing features at the end of each cycle. Thus, the releases are planned ahead of time and can be revised at the beginning of each sprint.
* Fourth is the “Sprint Planning” phase. Before each sprint, stakeholders hold a meeting to determine workload and methods of action.
* Fifth is the "Daily Stand-ups" phase. Here, there are short meetings at the beginning of each day that highlight each person's previous day's accomplishments and planned work for the day.
* Sixth is the "Sprint Review and Retrospective” phase. At the end of each sprint, the team holds a sprint review meeting to show stakeholders the finished work. Another meeting is held to discuss everyone’s performance during the sprint and changes to the next sprint that can avoid any problems.
* These are the phases of the Agile methodology that are based on iterative sprints and constant communication with the client. Changes are welcomed in this project management methodology.

### **Funding Requirements**

|  |  |  |
| --- | --- | --- |
| **Item** | **Details** | **Cost** |
| Data for our tool to learn from | The fish data that our proposed solution will use to train itself to predict correctly. | $0 |
| Upkeep servers that train the machine learning solution. | The machine learning solution will require a lot of time to train using various types of data. | $3,350 |
| New staff | The implementation of our solution will require the hiring of functional experts in machine learning. | $300,000 |
| New hardware and software licenses for staff | The new staff will require new hardware and software licenses for mandatory company work. | $23,000 |
| Server rooms for fish warehouses | The machine learning solution will be stored on-site so our warehouses will need to accommodate the new hardware. | $90,000 |
| Industry experts who can train all current staff in using the new solution | Our current employees will have to learn how to use this solution to reap all benefits | $120,000 |
| **Total Cost** | | $536,350 |

### **Data Precautions**

Sensitive or Protected Data:

* Our solution will not be using any sensitive or protected data. Thus, there are no general guidelines to review nor any necessary protocols required.
* The fish data that is being utilized has been sourced from Kaggle under a GNU General Public License. This public dataset remains freely usable by anyone.

### **Developer’s Expertise**

Developer Qualifications:

* Having studied machine learning and software development in my undergraduate studies, I have developed a strong ability to complete this task. My experience both inside and outside the classroom has prepared me for this project. My 2.5 years of internship experience have developed my ability to work in a professional environment. The completion of my bachelor's degree will be soon and having completed most of my coursework, I believe I had sufficient knowledge to accomplish this task.
* My qualification includes my future Bachelor of Science degree in Computer Science.

# Part B: Project Proposal

## Problem Statement

Problem:

* The problem is that due to the unstable motion of fishing boats, fishers are unable to weigh each fish individually. The sea’s waves prevent accurate weight measurements. The company’s current method of weighing a fish is to let it dry and then place it on a scale. This proves to be inefficient because the shelf life of the fish is lowered when waiting for it to dry and also during the weighing process.

## Customer Summary

Client/Customers:

* The project is being funded and developed by TastyFish and the client for this project is also TastyFish. This is because the project is aimed at improving the internal operations of TastyFish. Essentially this project is by TastyFish and for TastyFish.
* The clients that TastyFish serves include seafood distributors. Our company works with bulk distributors and not individual customers. Thus, our clients are the big corporations that deal with moving and selling fish.

Solving of Problem:

* This project will resolve our company’s core business problem successfully because it strives to improve one of the biggest downfalls of the company’s process. TastyFish relies on its commercial fishing operations to fund the company. The processing of the fish afterward is the most difficult part of the entire operation. The business problem focuses on the need to deliver a product that arrives to our clients faster and in better quality. TastyFish needs to outshine other competitors who have already implemented the newest technology in their fish harvesting processes. Thus, by potentially implementing machine learning into our business model, we would be able to boost productivity and throughput time. Our fish quality would increase as well, leading to a more valuable product for TastyFish to sell.

## Existing System Analysis

Current Tools:

* TastyFish currently does not use many applications or tools from modern times. The employees at the headquarters use basic laptops that connect to the company’s local servers. Windows 7 and Mozilla Firefox are used for most work. Gmail is used for communication. Excel is used for database storage. The fish measurements are gathered through handwritten documents and employees are tasked with transforming the data into CSV files.
* There are countless shortcomings in this current technological environment. The operating system is outdated and much of the information is recorded on paper before being converted into a digital form. The most shocking fact is that the company uses shared Excel files to collect and store data. Without a relational database, the company struggles to maintain modern data standards.
* Our solution is needed because, in an already slow and dying environment, the machine learning tool will drastically improve the throughput of fish products. Due to TastyFish’s avoidance of modern technologies, the company must start modernizing somewhere in its operations. A complete hardware and software overhaul will be extremely difficult considering the 200 locations that TastyFish has. In addition, this machine learning solution is easily transferrable to all platforms and devices. Thus, TastyFish will not have to worry about modernizing or modifying our solution.

## Data

Raw Data:

* The raw data is a CSV file downloaded from Kaggle. It includes seven columns of data points. These columns include the fish species, weight, and five length measurements. There are over 100 entries of data.

Data Handling:

* The data will be collected through a download link from Kaggle. The data will be processed by eliminating unnecessary columns. The data will be managed by ensuring the application has access to the data to train the model. The data will remain in the same directory as the application.
* During the development stage of the application development life cycle, the data will be collected and processed to train the linear regression model of our solution’s machine learning model.
* During the maintenance stage of the application development life cycle, the data will remain in the same file directory as the main Python application to ensure the Python file has access to the CSV file.

Data Anomalies:

* If necessary, any erroneous rows will be deleted from the CSV data that holds the raw data.

## Project Methodology

Agile Development:

* The Agile project management methodology is great for solutions that require iterative work like ours. Constantly having to train a machine learning model requires iteration which would align with the project steps. In addition, having to work with TastyFish to approve the user interface will bring about many changes even after the designs are published.
* Because our user interface will use a console to relay text, it will not be too difficult to respond to TastyFish’s change requests. If the front end of our application had been developed using a complex stack including Angular or jQuery, it would have been troublesome to use the Agile methodology.

Development Plan Using Agile:

* As mentioned earlier, we will be using a reference to Adobe’s online article to standardize the step we will take using Agile methodology.
* First is the “Project Planning” phase. We will discuss the goals and scope of the project here. Our initial goals and scopes should include machine learning but any mention of linear regression or Python’s “scikit-learn” library is unnecessary.
* Second is the “Product Roadmap Creation” phase. The backlog will be created using the various deliverables that are decided. The primary deliverables that will be listed include the console user interface, the functional Python files, etc.
* Third is the “Release Planning” phase. Here, we will plan on the features that will be released for the various sprints. Features such as those involving the "sklearn.metrics” module will be grouped to ensure relevant development occurs simultaneously
* Fourth is the “Sprint Planning” phase. The work and method of action for each sprint will be decided here. It is integral that each spring planning includes a metric to track the accuracy of the linear regression model. It is also important that any planned developments receive a unique branch in the GitHub repository.
* Fifth is the “Daily Stand-ups” phase. The daily meetings that explain each contributor's work and their upcoming obstacles will be necessary. With new hires, some of the software that our company uses will need to be taught to the new hires. Various documentation on our company's internal code reviews will be needed to ensure standups remain relevant to the sprint and not the new hires' onboarding troubles.
* Sixth is the "Sprint Review and Retrospective” phase. At the end of each sprint, the team will review the work and any necessary improvements to the next sprint. All pull requests in the GitHub repository must be accepted and merged into the main branch.
* This outline details the necessary phases. All important details like timelines, contracts, and deliverables will be assessed in the project planning phase.

## Project Outcomes

Deliverables

* The user guide will be a deliverable that is expected to be worked on during and after the creation of the project. It will serve as a set of instructions for any basic user to install and use the application.
* There will be three visuals that depict the various descriptive and non-descriptive methods. They will illustrate the data. They will be deliverables because they will help TastyFish’s internal IT department to visualize the data and final algorithm.
* The completed application will be a deliverable that will serve to provide all the functionality mentioned above. These functionalities include the tool that will predict the weight of the fish using a linear regression model.
* The trained linear regression model will be a deliverable that must be completed and delivered to TastyFish. This deliverable will be crucial as it will mark the successful implementation of machine learning into the finished application.
* The processed data set will be a deliverable as it will be required to train the linear regression model.

## Implementation Plan

Implementation Outline

* The general strategy for implementing the project is to approach the project with a mindset that is focused on optimizing time and effort. Because many project tasks can be completed simultaneously, the timeline must be optimized. Tasks like creating the machine learning model, training the machine learning model, and creating the user interface are various project milestones that will be worked on simultaneously.
* The phases of the project rollout will be determined by the sprints from the Agile methodology. Due to our project needing constant client feedback through iterative changes, product releases will be incremental and done bi-weekly.
* The dependencies of the project will primarily focus on the Python libraries and the data set. Both are public domain so it will not present any issues when attempting to integrate these dependencies into the project.
* Testing of the machine learning algorithm will include unit tests that will be carried out by the development team. Ad hoc testing will be utilized sporadically throughout the project to ensure the model is following a good development path. The development of the project will be focused on white box testing due to the complex nature of the linear regression model. The distribution of the application will be planned after the model reaches sufficient accuracy.

## Evaluation Plan

Verification Method:

* During each stage of development, peer reviews will be conducted as the project’s standard verification method. The peer reviews will allow the team to review the code of the other team members. Since the team will consist of newly hired functional experts in machine learning, the verification of each development stage will serve as a sufficient means of quality control and assurance.

Validation Method:

* After the project, a complete white box testing process will be utilized. The entire team will peer review the final product and each member will utilize white box testing to decide whether the final product is worthy of being released. The functional experts will spend the majority of their time evaluating the efficiency and accuracy of the linear regression model.

## Resources and Costs

Labor Time & Costs:

* There will be hiring of two different staff, as specified in the “Funding Requirements” section of part A. The first type of staff will cost $300,000. This staff will be responsible for helping implement the project solution and also will be functional experts in machine learning. The starting salary for these positions will be advertised as $95,000. Thus, there will be three new functional experts to be hired. The remaining $15,000 will be utilized as a buffer for any negotiations.
* The second type of staff will cost $120,000. This staff will be responsible for training the company in the usage of the new project solution. The starting salary will be advertised at $60,000 for each position. There will be no negotiations allowed. The two staff that will be hired will be responsible for holding courses for current staff to learn about the new solution.
* These salaries are budgeted according to one year's worth of salaries. The estimated project duration will be one year. Thus, the employment of these five new staff members will be reevaluated after the year.
* These employees will be working as full-time employees and are exempt from overtime pay. They will work the normal 40 hours per week. No overtime will be expected nor allowed from these employees.

Hardware & Software Costs:

* The data set that contains the training information will be free as it is provided under a free license by Kaggle.
* The five new staff will each be provided software licenses for our company's basic work tools. They will also be provided with a variety of new hardware, including laptops, computers, smartphones, etc. The new Microsoft Office licenses will cost $500 in total. The new Windows 10 licenses will cost $500 in total. The smartphones will be provided for each staff and cost a total of $3,500. The powerful computers provided will cost $15,000 in total. The new laptops will cost $3,000 in total. The new webcams, microphones, and Bluetooth headsets will cost $500 in total. The total will come out to $23,000, which is the same as listed in the aforementioned budget graph.

Environment Costs:

* The new server rooms that will host the machine learning solution will be quite expensive because they will be expected to hold terabytes of fish data and also be fast enough to return predicted weights in a short time. Only twenty out of the total two hundred TastyFish locations handles enough fish to make the machine learning solution worth the cost. Thus, using the total $90,000 budget for server rooms, the twenty locations will split the budget for a total of $4,500 each. The processors will be the most expensive running around $800 each. The memory will total around $300 each. The storage solution will consist of countless SSDs for each location, costing around $1,000. The remaining $2,100 will be used for the construction of the server room and temperature control.
* Finally, the upkeep of the servers will cost $3,350 in total. These factor in the electricity costs and the monthly subscription software licenses. The subscriptions for the monthly licenses include VPN and cloud storage solutions.

## 

## Timeline and Milestones

|  |  |  |  |
| --- | --- | --- | --- |
| **Start Date** | **End Date** | **Duration** | **Milestone** |
| 09/01/2023 | 10/30/2023 | 2 months | Project planning will occur. All necessary resources, budgets, contracts, staff, hardware, software, and tool will be collected. |
| 11/01/2023 | 04/30/2024 | 6 months | The development will begin. All coding, development, and model training will begin. |
| 05/01/2024 | 05/30/2024 | 1 month | Testing of application begins. Peer reviews will be conducted. |
| 06/01/2024 | 06/30/2024 | 1 month | Revisions based on testing and feedback. |
| 07/01/2024 | 07/30/2024 | 1 month | Server rooms will be constructed and set up according to application specifications. |
| 08/01/2024 | 08/30/2024 | 1 month | Complete rollout of application to the designated locations. |

# Part C: Application

Submitted Files:

* The project files do not require any links. All necessary files are located in the fish-weight-estimation.zip file. After unzipping the file, the folder will contain all project files, documentation, writeups, visuals, etc.
* The two files necessary to run the application are the main.py file and the fish\_data\_processed.csv file. The environment and additional software will be the responsibility of the user. Directions will be provided in the “User Guide” section of Part D.

Components of the Project:

* The project contains three visualizations that are located in this document, in the visuals folder, and also inside the application. The application can generate the three visuals.
* The three descriptive methods in the project include the histogram, the scatter plot, and the graph displaying the prediction error of the regression model.
* The one non-descriptive method in the project is the linear regression algorithm that is included in the project’s code in the main.py file.
* The application of machine learning is seen through the non-descriptive method. The linear regression model is trained using the data set to improve its ability to predict the weight of a fish using its dimensions.
* An interactive dashboard is provided through the console in the IDE. Using text, the user interface mimics a navigatable menu using commands such as “A”, “B”, and “end”. The user can interact with the machine learning model by inputting fish dimensions to receive a predicted fish weight.
* The user interface is user-friendly because the “User Guide” section from Part D explains 23 steps to installing and using the application. There are images provided in almost all 23 steps to provide further clarification. There are links provided that give information about the software used in situations requiring troubleshooting.
* The application can run using only two files: the CSV file and the Python file. Because of this, all operations are done locally and do not require a constant network connection. The offline nature of the application protects it from external attacks because it will never require any contact with an outside source after it has been installed on a device. Thus, the security of the application remains strong.

# Part D: Post-implementation Report

## A Business (or Organization) Vision

Description of Company:

* TastyFish is a company that catches and sells fish for consumption. Due to laws and regulations for commercial fishing, the fishers are required to measure various lengths of each fish they catch. If the dimensions fall below a certain threshold, the fish must be released to ensure the younger population can breed.

The Problem:

* The problem was that due to the unstable motion of fishing boats, fishers were unable to weigh each fish individually. The sea’s waves prevented accurate weight measurements. The company’s previous method of weighing fish was letting it dry and then placing it on a scale. This proved to be inefficient because the shelf life of the fish was lowered when waiting for it to dry and also during the weighing process.

How the Application Solved the Problem:

* The application solved the problem by using machine learning to predict the weight of each fish. A linear regression algorithm was applied to the machine learning model and by training the model on existing fish data, the model was able to learn how to predict the weight of a fish given its vertical length, diagonal length, cross length, height, and diagonal width.

Use of Application to Solve the Problem:

* A user can use the application to predict the weight of a fish without weighing it. If the user possesses the five dimensions of the fish, the user can input those dimensions into the trained linear regression model to predict the weight of the fish in grams. An example would be if the user inputted the string “23.2, 25.4, 30, 11.52, 4.02”, then the model would provide a predicted weight of 325.23 grams.

Screenshot of Application Console Output

Graphical user interface, text, application

Description automatically generated

## Datasets

The Raw and Processed Data

* The raw data is a CSV file that contains seven columns. The columns consist of the fish species, weight, vertical length, diagonal length, cross length, height, and diagonal width.
* The processed data is a CSV file that contains six columns. The columns are the same as the raw data except for the first column is removed.

Processing of Raw Data

* The raw data were processed by removing the first column for fish species. Little processing was needed because the raw data was minimalistic and contained mostly necessary columns. The data in its raw form was already accessible to the algorithm but processing the raw data removed the need for the data frame object to iterate over an unused index in the array. Thus, a minuscule amount of processing power was saved by removing the first index of the arrays in the CSV file. The algorithm would have returned the same results if the raw data was used with the appropriate array indexes changed.

Examples of the Raw and Processed Data

* Below are the first three rows from the raw data file named "fish\_data\_raw.csv"
  + Species,Weight,Length1,Length2,Length3,Height,Width
  + Bream,242,23.2,25.4,30,11.52,4.02
  + Bream,290,24,26.3,31.2,12.48,4.3056
* Below are the first three rows from the processes data file named "fish\_data\_processed.csv"
  + Weight,Length1,Length2,Length3,Height,Width
  + 242,23.2,25.4,30,11.52,4.02
  + 290,24,26.3,31.2,12.48,4.3056

Access to Datasets:

* The original dataset is available for download from this link:
  + <https://www.kaggle.com/datasets/aungpyaeap/fish-market>
* The dataset contained a single CSV file.
* The “fish\_data\_raw.csv” file is the exact dataset that was downloaded from the link.
* The “fish\_data\_processed.csv” file is the only dataset that was used throughout the project.

## Data Product Code

Review of Code Functionality:

* In the analysis and development of the application, the code was used to execute a variety of functions.
* Overall, the product does exactly as it is supposed to. It takes in five numbers that represent various measurements of a fish and predicts the weight of the fish using a trained linear regression model.
* Because the raw data was preprocessed manually, the code never touched the raw data. Only the processed data was touched by the code. This is because the processing only involved the deletion of one column in a CSV file. This saved the code from having to iterate over that one column, which is an insignificant saving of processing power. The processing of data allowed the code to use all columns of the data.
* The code provided visualizations of data using imported libraries. These visualizations were localized to the “visuals()" function in the "main.py" file. The project used the "matplotlib" library to allow all three visualizations to appear on the screen in individual windows. These three visualizations were the histogram, scatter plot, and graph displaying the prediction error of the regression model. The "metrics" module from the "scikit-learn" library was used to create the prediction error graph. The "pandas" library was used to create the histogram and scatter plot.
* The code aided in the development of the descriptive methods because as mentioned earlier, the histogram, scatter plot, and graph displaying the prediction error of the regression model, were all generated using Python libraries. A descriptive method is anything that describes data. The histogram described the distribution of each variable in the CSV file. The data showed that there were certain numbers more prevalent than others in each variable. The scatter plot showed the correlation of each variable in the CSV file. As each variable's measurement increased, the scatter plot showed a positive correlation with all other measurements. The diagonal across the scatter plot displayed the histogram data. The graph displaying the prediction error of the regression model showed the difference between the predicted weight values and the actual weight values. It showed that the extreme ends of the weight predictions were more prone to error than the predictions of weight values near 200 to 700 grams.
* The code aided in the development of the non-descriptive methods because the “scikit-learn” library provided many machine learning modules to train a linear regression model. The “pandas” library was used to create the data structure for the model. After that, the “scikit-learn” library had modules to create linear regression models, train the models with data, and provide predictions with new data.
* The data analysis done through the visuals and reviewing of the CSV file helped created correlations between the independent variables that lead to the prediction of the dependent variable. Seeing that there was a positive correlation between the independent variables and the dependent variable helped choose the scatter plot because the data points allowed one to realize that as the independent variables increased, the dependent variable increased as well.
* The analysis of the data leads to the creation and improvement of the histogram. Because of the data reviews, it was seen that the model could predict values more accurately when there was more existing data numerically similar to an input.
* This leads to the creation and improvement of the graph that displays the prediction error of the regression model. Due to the histogram showing trends in the existing data, the prediction error followed the histogram. This means that as the variables increased in quantity around a certain numerical range, the prediction error decreased. This proves the use of the prediction error in showing that the histogram data was relevant in predicting the accuracy of the model.

Non-Descriptive Method:

* The non-descriptive method of the project was the machine learning algorithm. The project used a supervised learning algorithm. Instead of a classification algorithm, a regression algorithm was utilized because a linear regression algorithm was used to predict the weights.
* The linear regression model was developed by using a Python library called “scikit-learn”. It was also improved using methods within that Python library.
* This use is justified because the independent variables and one dependent variable were all numerical. With the five independent variables and one dependent variable, a numerical prediction called for the use of the linear algorithm. Because it was numerical and not qualitative, a logistic regression algorithm was not used as the prediction was not a binary choice.
* The linear regression model was trained using the dataset and also tested using the dataset. The “test\_size” variable was set to 0.33 which meant ~33% of the dataset was used to train the model and ~66% of the dataset was used to test the model. Using the existing variable data, the model was trained to predict the weight of the fish given five other measurements.
* The training process was appropriate because the default “test\_size" variable was 0.25 and in any machine learning model, you need data to both train and test the prediction algorithm. Thus, the process of the algorithm development followed a reasonable plan with adequate justification.

The “test\_size” Variable



Project Source Code:

* Because the project is a standalone app that is run in an IDE, the project only requires the "main.py" file and the "fish\_data\_processed.csv" file to provide full functionality. The submission of this project will include these files, which should be run in an IDE (ideally PyCharm).

The 2 Most Important Project Files



## Objective (or Hypothesis) Verification

Project Objective:

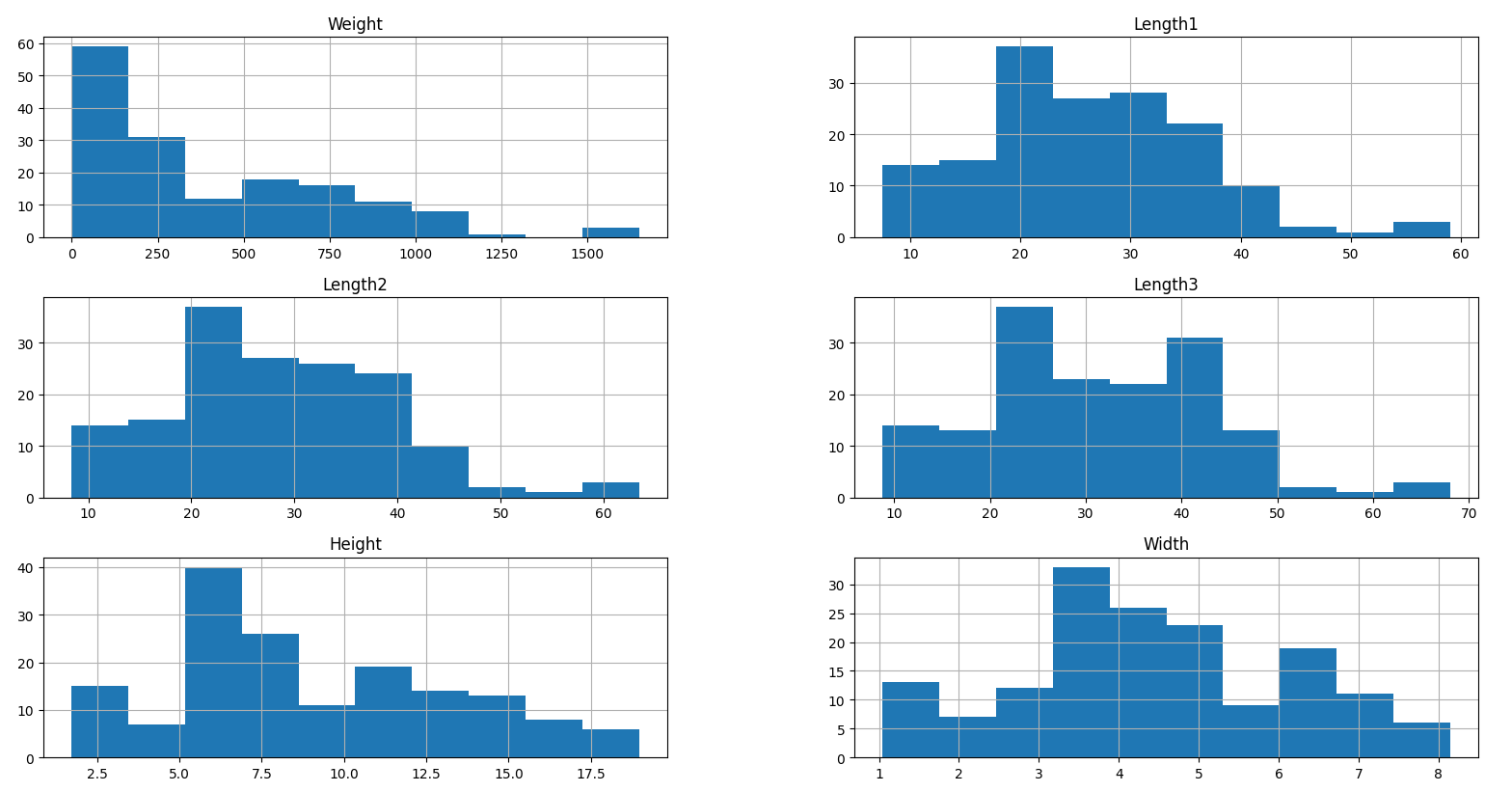
* The objective of the project was to create a machine learning model that would learn from existing fish data and be able to predict the weight of a fish given its five independent measurement variables.
* The objective was met because the resulting project was a linear regression model that used supervised learning to train from the fish dataset. As of April 5, 2023, the R-squared value of the linear regression model was 0.91. Thus, the linear regression model can predict the weight of a fish with some margin of error.
* “R-Squared (R² or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model (the goodness of fit)” (Taylor).

## Effective Visualization and Reporting

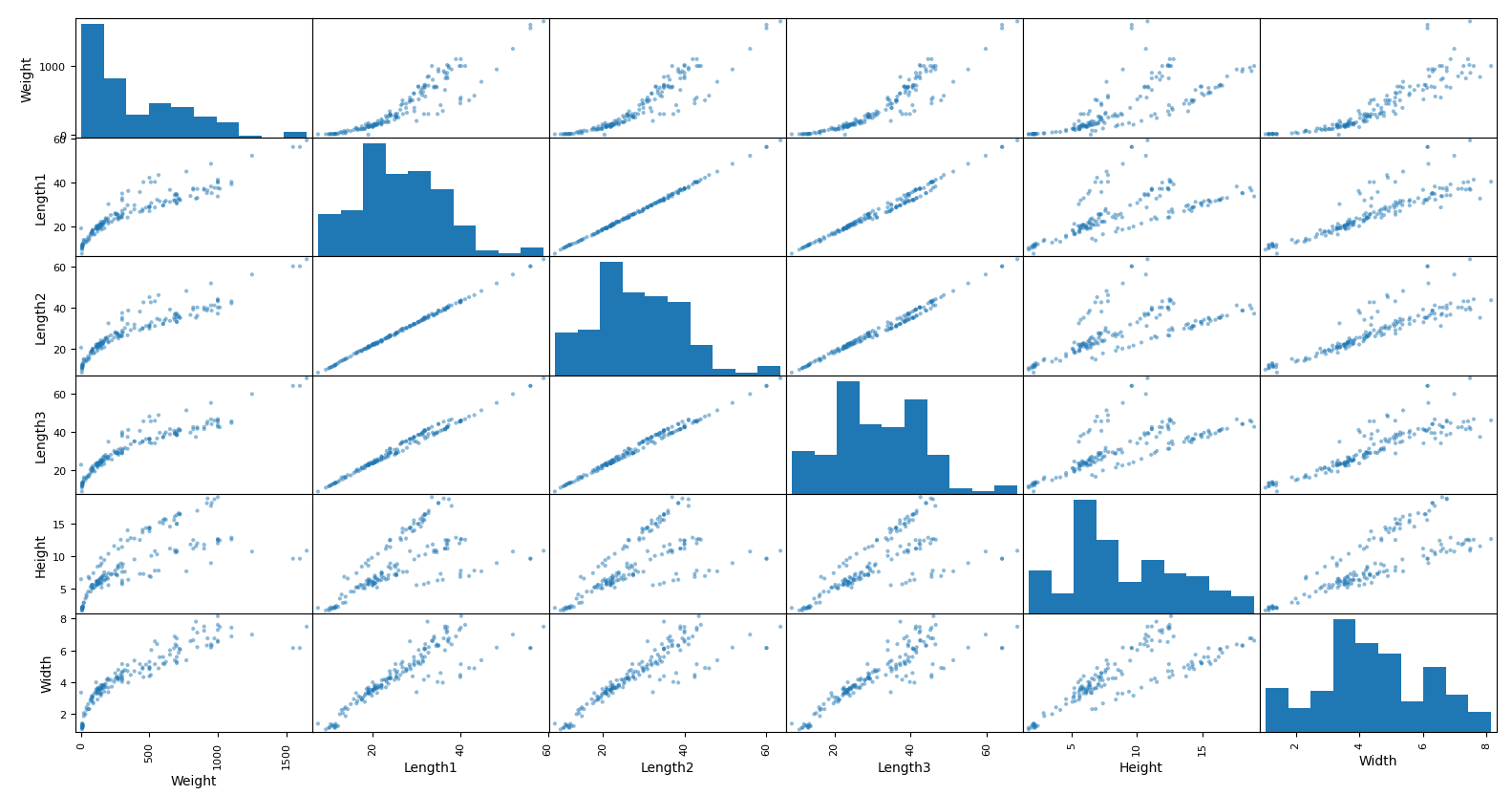
Non-Descriptive Development Process:

* The development process behind creating and training the linear regression model was supported by visualizations and descriptive methods. At the most basic level, the development process by first exploring the data. The dataset was explored using various visualization tools to identify trends. Once it was observed that the weight and all five other measurements of the fish has a positive correlation, a linear regression model was seen as an appropriate approach. During the data exploration, it was clear that the five independent variables positively influenced the dependent variable.
* During the analysis of the data, it was clear that the linear regression model would be better suited when using measurements that were close to the mode (highest frequency) of the data. This helped when developing the non-descriptive method (the linear regression model) because it identified that although the R-squared value might be 0.91, there was a higher prediction error in the upper and lower ranges of the weight predictions.
* The data summaries included the three visualizations and the R-squared value. These four data summaries provided a better insight into how the linear regression model would perform in certain situations.
* The three visualizations (the histogram, scatter plot, and graph displaying prediction error) are shown below. For a clearer image, refer to the “visuals” folder in the project directory for the image files.

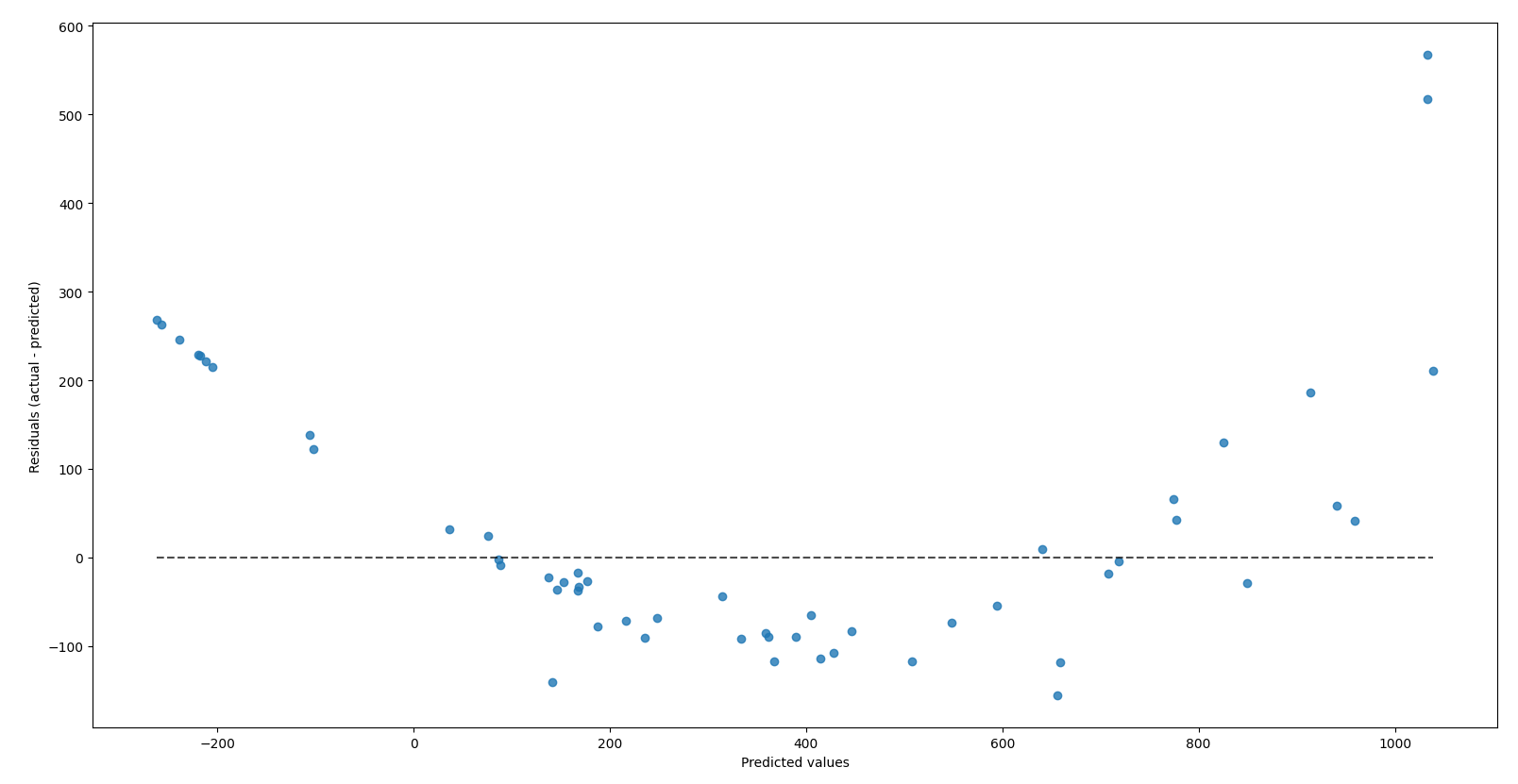
Histograms



Scatter Plots



Prediction Error of Regression Model



## Accuracy Analysis

The Model’s Accuracy Metric:

* The metric used to assess the accuracy of the linear regression model was an R-squared value.
* “R-Squared (R² or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model (the goodness of fit)” (Taylor).
* The R-squared value is a decimal number that ranges from 0.0 to 1.0. Our product returned an R-squared value of 0.91.
* Below is an example of the R-square value calculation in the project code. The calculation began with the non-descriptive method (linear regression model) predicting the weights of the fish using the test data of five measurements. Then, the predicted weight values were compared against the correct weight values. This variance was the R-squared value and illustrated the variance in the regression model versus the actual data.
* This metric was very accurate in proving the linear regression model’s accuracy because it calculated the variance of the true data against the predicted data. When dealing with numerical variables like in this project, it was the only way to display how accurate the regression model was.

Code Screenshot: Calculation of R-squared Value

Text, letter

Description automatically generated

## Application Testing

Testing Methods, Process, and Results:

* The application was tested only by me. The testing process included using a variety of valid and invalid inputs for the linear regression algorithm. In addition, the menu navigation in the console was also tested to account for various errors and situations.
* The results of the testing were significant because they led to many changes in the code that improved the final product. The first test involved inputting various combinations of the five numerical inputs for the linear regression algorithm. The outputs were compared to the actual data to identify similarities and to ensure that the final prediction was not outlandishly inaccurate.
* The testing of the application's user interface also brought about many changes. By testing the application from the end user's perspective. The graphical interface was altered to improve visibility. The print statements were altered to include newline operators. If the linear regression algorithm was fed an invalid input, the appropriate error message was given. By testing the variety of inputs, the code was able to improve by including "try-except" statements. These prevented errors from the algorithm receiving too few or too many numbers. In addition, if any of the inputs included a non-numerical value, a corresponding error was displayed.
* The use of the "A", "B", and "end" inputs was added after testing showed that it was not possible to maneuver between the reporting section and the calculation section of the application. This led to the improvement of the code where the aforementioned three strings allowed the user to navigate between the program’s different functionalities without having to terminate the program and start over again.
* Various parts of the application displayed error messages if the program was terminated while waiting for input. To prevent this, a "try-except" statement included a "KeyboardInterrupt” error case where the program would quit instead of displaying an unnecessary error.

## Application Files

Location of Files:

* The 8 files mentioned below will all be located in the provided “zip” file. All files are located directly inside the directory.

Required Files:

* The project requires only two files to execute properly. Because the entire project and associated documentation are located inside the project folder, everything mentioned in this section can be accessed through the provided "zip" folder. The necessary files to execute the application are the following:
  + main.py (run this file in PyCharm IDE to access the console UI)
  + fish\_data\_processed.csv (place this file with “main.py” for regression model training)

Not Required Files:

* There are supplementary files that are not necessary to access the application. These files only serve to improve the viewer’s understanding of the project. They are not needed to execute the program. These files are the ones mentioned below:
  + “visuals” folder
  + task\_1\_writeup.docx
  + task\_2\_documentation.docx
  + fish\_data\_raw.csv
  + README.md

## User Guide

Installation and Use Guide:

* If any of the following software is already installed and configured, you can skip that step.
* The following directions apply to a Windows 10 machine.
  + - 1. Download the fish-weight-estimation.zip file onto your computer.



* + - 1. Unzip the fish-weight-estimation.zip file. You can use this link (<https://www.7-zip.org>) to install a file unarchive tool to unzip it. Refer to this link (<https://www.7-zip.org/faq.html>) if any issues arise during the installation of 7-Zip.

Shape

Description automatically generated with medium confidence

* + - 1. Download and install PyCharm Community Edition from this link (<https://www.jetbrains.com/pycharm/download>). Refer to this link (<https://www.jetbrains.com/help/pycharm/installation-guide.html>) if any issues arise during the installation of PyCharm.

Graphical user interface, text, application

Description automatically generated

* + - 1. Download and install Python from this link (<https://www.python.org/downloads>). Refer to this link (<https://wiki.python.org/moin/BeginnersGuide/Download>) if any issues arise during the installation of Python.

Graphical user interface

Description automatically generated

* + - 1. Open PyCharm.



* + - 1. In PyCharm, go to File -> Open . . . -> select the unzipped fish-weight-estimation folder. Click “OK”.

Graphical user interface, application

Description automatically generated

* + - 1. With the project opened in PyCharm, open the main.py file in the IDE.

Graphical user interface, text, application

Description automatically generated

* + - 1. In PyCharm, configure the Python interpreter for this project using the Python installation from step 4. Use this link (<https://www.jetbrains.com/help/pycharm/configuring-python-interpreter.html>) if any issues arise during interpreter configuration.

Graphical user interface, table

Description automatically generated

* + - 1. In the main.py file, ensure lines 1 – 4 are not showing errors. If they are showing errors, it means that the required Python libraries are not installed. On each error line, right-click the problematic import and click “Show Context Actions”. Choose the option to install the corresponding Python library. Repeat this for all four lines of imports

Graphical user interface, text, application

Description automatically generated

* + - 1. After the libraries are installed, run the main.py file.

Text

Description automatically generated with medium confidence

* + - 1. The console should appear now. In the console, type in “B” and press ENTER.

Graphical user interface

Description automatically generated with medium confidence

* + - 1. The console will generate the R-squared value of the linear regression model.

A picture containing chart

Description automatically generated

* + - 1. Three windows will appear showing the prediction error of the regression model, a histogram, and a scatter plot. These three visuals will be discussed in this document. Enlarged versions are available in the “visuals” folder and also in the “Effective Visualization and Reporting” section of Part D.

Chart

Description automatically generatedChart

Description automatically generatedChart, scatter chart

Description automatically generated

* + - 1. After closing all three windows, the console will automatically return to the main menu.

Graphical user interface, text, application

Description automatically generated

* + - 1. In the console, type in “A” and press ENTER.

Graphical user interface, text, application, Word

Description automatically generated

* + - 1. The console will generate the menu to use the trained linear regression model that predicts the weight of a fish given five numbers (the vertical length, diagonal length, cross length, height, and diagonal width).

Graphical user interface, text, application, email

Description automatically generated

* + - 1. In the console, type in “23.2, 25.4, 30, 11.52, 4.02” and press ENTER to test the model.

Graphical user interface, text, application

Description automatically generated

* + - 1. The output will return the predicted weight of the fish in grams.

Graphical user interface, text, application, email

Description automatically generated

* + - 1. Use this menu to repeatedly test variations of fish measurements.
      2. When want to close the program, type “end” in the console and press ENTER.

Text

Description automatically generated

* + - 1. Again, in the console type in “end” and press ENTER.

Graphical user interface, text, application, Word

Description automatically generated

* + - 1. The program is now terminated.
      2. Return to step 10 to start the program again.

## Summation of Learning Experience

Application of Previous Knowledge and Experience:

* This project was made easier because of my previous experience in Western Governors University’s (WGU) coursework. During my experience earning my Bachelor of Science in Computer Science degree, I learned many hard and soft skills. I learned how to better use various IDEs like PyCharm and IntelliJ. The practical experience using these IDEs in previous projects brought me knowledge of various aspects. Things like configuring the project's interpreter, installing the appropriate compilers, and creating a proper project structure were all things that I learned through WGU.
* The use of version control using Git was also integral in the development of this project. Ensuring that there were safe backups of my work in the cloud brought an extra layer of confidence to my work. In addition, the ability to roll back any changes brought comfort when changing and improving my code. WGU helped me develop my skills in Git because of the many projects that need to be developed during the coursework. In addition, working with virtual machines in previous projects required me to be competent in Git.
* Previous projects that included developing code and solving a business need helped me understand the steps I needed to take to solve a solution with software. Using Python in this project was easy since I already had experience using Java in courses like “Software I” and “Software II”.

Application of New Knowledge:

* Additional knowledge was needed to complete this project. The topic of machine learning was explored in my previous courses but never practically used to the extent of this project.
* I needed to learn various topics like supervised learning, unsupervised learning, reinforced learning, regression algorithms, and classification algorithms. These topics of machine learning were important in understanding the final product and how I was going to achieve it.
* I also needed to learn how to use various Python libraries such as the “scikit-learn” library to implement linear regression into my data. I needed to learn how to use the “pandas” library to implement “data frame" structures that help CSV data. I also needed to learn the "matplotlib” library to be able to create visuals from my data and model.
* I learned more about the application of visuals in data sets and regression models to create the visuals.
* I needed to expand my knowledge about the”try-except" statement and the keyboard input error when creating my user interface.
* I needed to learn about descriptive methods and non-descriptive methods so that I could properly implement a way to describe my data and also infer new knowledge from it.

Project Contribution to Lifelong Learning:

* My definition of lifelong learning is “the continual pursuit of new knowledge that will improve one’s expertise in a subject matter”.
* When I analyze this experience for its contributions to my lifelong learning, I believe it has opened up a new pathway for me to develop a better understanding of machine learning. Before this project, I had no ambition for machine learning and the concept was very foreign to me. However, this project has caused me to become interested in the topic. This will lead me to pursue more opportunities and learning moments where I can expand my knowledge of machine learning.

# Part E: Sources

Adobe Communications Team. “Beginner’s Guide to Agile Project Management”. Adobe Experience

Cloud Blog, 18 Mar. 2022, https://business.adobe.com/blog/basics/agile#what-are-the-6-steps-in-

agile-project-management.

Taylor, Sebastion. “R-Squared”. Corporate Finance Institute, 4 Mar. 2023,

https://corporatefinanceinstitute.com/resources/data-science/r-squared.