**Capstone: Bike Store Machine Learning Algorithm**

Danny Nguyen

011014337

Western Governors University

Computer Science Capstone C964

June 8th, 2023

C964: Computer Science Capstone

Task 2 parts A, B, and D

[Part A: Project Proposal for Business Executives 3](#_Toc1515453419)

[Letter of Transmittal 3](#_Toc1044087116)

[Project Recommendation 4](#_Toc1721750458)

[Problem Summary 4](#_Toc205640471)

[Application Benefits 4](#_Toc739883765)

[Application Description 5](#_Toc1967641075)

[Data Description 6](#_Toc1580263866)

[Objectives and Hypothesis 6](#_Toc1084915425)

[Methodology 7](#_Toc1455093807)

[Funding Requirements 8](#_Toc1716359054)

[Data Precautions 8](#_Toc821607173)

[Developer’s Expertise 8](#_Toc742297278)

[Part B: Project Proposal 8](#_Toc1516064891)

[Problem Statement 8](#_Toc602847551)

[Customer Summary 9](#_Toc411678477)

[Existing System Analysis 9](#_Toc1668713332)

[Data 10](#_Toc1911399030)

[Project Methodology 10](#_Toc1307313575)

[Project Outcomes 11](#_Toc1967172832)

[Implementation Plan 12](#_Toc306674583)

[Evaluation Plan 13](#_Toc849271423)

[Resources and Costs 13](#_Toc1695791095)

[Timeline and Milestones 14](#_Toc787992355)

[Part C: Application 15](#_Toc1767968421)

[Part D: Post-implementation Report 16](#_Toc22579049)

[A Business (or Organization) Vision 16](#_Toc1496854501)

[Datasets 16](#_Toc645781746)

[Data Product Code 18](#_Toc157285634)

[Objective (or Hypothesis) Verification 20](#_Toc1315546441)

[Effective Visualization and Reporting 21](#_Toc656892821)

[Accuracy Analysis 23](#_Toc1956352798)

[Application Testing 24](#_Toc871280264)

[Application Files 25](#_Toc1639588139)

[User Guide 25](#_Toc1125023771)

[Summation of Learning Experience 31](#_Toc1702287338)

# Part A: Project Proposal for Business Executives

## Letter of Transmittal

Mr. Odenkirk,

Bob’s Bike Shop

555 Bill Street

Houston, TX 11111

Hello Mr. Odenkirk,

I am writing to your organization because I believe I can solve a business problem that you guys have. As you may know, there are many concerns regarding the supply of classic jersey sizes and how it relates to the customer demographic. Right now, it’s difficult to tell which demographic is buying which classic jersey size, and the manufacturers are either over producing certain sizes or underproducing other sizes. This is leading to a inefficient budget allocation and it needs to be changed.

Fortunately, you have many records of sales that can be utilized to determine the customer demographic and the most likely classic jersey sizes they will buy. I believe that if you can use the power of machine learning algorithms, you’ll be able to gain better insight of your customer demographic. I will use bar charts and other data analysis tools to gain a better understanding of where your customer base lies and create a machine learning algorithm that predicts the size a customer will purchase based on their age and their gender.

The problem that the machine learning model will solve is that it will provide better insight to what a specific customer demographic wants. Let’s say you wanted to know what certain size a 34 year old male would buy. You would simply input the data into the machine learning model and it will predict the size that they will most likely buy. By doing this, you can gain a better understanding of your target audience and therefore produce more of those T-shirt sizes to target those specific customers.

The estimated budget required to implement this strategy will be $12,000 with a maintainability of $2,000 a year. With my bachelors in computer science and my 10 years of experience in working as a data engineer and machine learning engineer, I am more than confident that this will be a valuable return on investment. Thank you for taking the time to read this letter.

Sincerely,

Danny Nguyen

## Project Recommendation

### Problem Summary

As Bob’s Bike Shop continues to grow in size, it’s important to gain a deeper understanding of customer insight and their preferences. Right now, there seems to be a discrepancy between the amount of jersey sizes being produced and the customer demand. We will need to find a way to predict what the customer wants and what resources are the most used. By doing so, we will gain better insight into what our customer base needs and therfore better allocate our resources.

### Application Benefits

The benefit that this application will provide is that it will give us better understanding of what jersey sizes each customer demographic will want. These data analysis tools will allow us to better re-allocate our resources and target certain demographics with certain sizes. As a result, less money would be wasted producing certain jersey sizes and resources would be used more efficiently.

### Application Description

let’s say we wanted to produce a certain jersey size to target a demographic of 34 year old males. We would simply input the data into the machine learning algorithm and it would predict the estimated shirt size that the specific demographic would most likely buy. Bar charts would also be used to describe the customer demographic, giving more insight on who our main target audience is. Lastly, a confusion matrix would be shown in order to show the accuracy of the model. From this new revelation, we can direct our marketing strategy towards that specific demographic and produce more of in-demand sizes.

### Data Description

The data that will be used to create the application will be from the “Bike Sales in Europe” dataset from Kaggle.com. The input data used in the model will consists of gender, a categorical data type, and age, a numerical data type. The output data measured is ordinal since it includes different jersey sizes, ranging from smallest to largest. The data structure used in this project will be a structured table, allowing for effective query filtering. The independent variables used in this model will be both the gender and the age of the target demographic. The output, or dependent variable would be the jersey size. The limitation that this application will have is that it does not account for other products that the customer may buy. The only variable measured will be the classic jersey sizes, not other products in inventory.

### Objectives and Hypothesis

The objective of this project is to be able to analyze and predict what jersey size each customer demographic is likely to buy. By using a machine learning algorithm built on previous data, we would be able to gather intel and produce more in-demand jersey sizes. The accuracy of the model will be determined by the data itself, so it’s important for us to also analyze and interpret the data. The goal for this project is to reach an accuracy of at least 75%. As more records of jersey sizes will be inputed into the data, the machine learning algorithm will become more accurate in its assessment. The goal of this project is not just the accuracy, but also implementing the correct visualization tools for data analysis and gathering intel.

### Methodology

The methodology used to create this project will be done using the waterfall method. The benefits of using the waterfall method to implement in this project is that it clearly defines the requirements and the end goals. It also allows for accurate budget requirements and it can be easily implemented in the early design stages. There are no obsurities and everything is clearly defined.

The first step is to filter out all the records that do not include the jersey models, this will allow for better predictions and remove unneccary training data that will not benefit the machine learning model. We will most likely use SQL Server Management Studio to query the necessary data. The second step is to implement the machine learning algorithm based on the inputted data. The user should be able to input any age and gender and the machine learning algorithm would try to predict the jersey size shirt they will buy. We will then add in bar charts that will visualize the main customer demographic to provide better insigjts. Lastly, we will create a confusion matrix that measures the accuracy of the machine learning algorithm.

1. Requirements: We will work with stakeholders, manufacturers, and dataset websites to gather the data. We will also begin filtering out necessary data for machine learning implementation.
2. Design: We will create a user friendly interface that is easy for users.
3. Implementation: We will add in the predictive machine learning model as well as visualization tools
4. Verification: We will add in the confusion matrix and accuracy test of the model.
5. Maintenance: Bugs will be fixed as needed.

### Funding Requirements

The machine learning algorithm and data used in this project will be free and open sourced. The estimated costs that the software development team will be $200 an hour consisting of 40 hours. This will bring a charge up to $8,000. An additional $4,000 will be used for debugging puprposes and hardware equipment. In order to maintain the software, $2,000 will be needed each year. This will be used to fix any upcoming bugs as well as updating the dataset for better accuracy. In sumamry, the total cost of this project will be $12,000 and $2,000 each year for maintainability.

### Data Precautions

Since kaggle.com is providing the dataset used for the machine learning algorithm, there is no sensitive data within the dataset. The only fields included in these datasets are the date of transaction, customer characteristics, products they bought, and the costs and revunue of the purchase. We will still filter out records by using SQL in order to make our machine learning algorithms more accurate in regards to jersey sizes, but we do not have to worry about sensitive data, like social security and credit card information.

### Developer’s Expertise

The developer leading this project has a degree in computer science and 10 years of working experience. They have developed many different machine learning models for other businesses and data analysis tools to provide better insight. With this unstoppable track record, I am more than confident that this developer will deliever the application effectively.

# Part B: Project Proposal

## Problem Statement

As you may have known, there is an increase descrency between the amount of jersey sizes being produced and the customer demand. We have either produced too much of one size, or not enough of another size. Bob’s Bike Shop requires a thorough analysis of which jersey sizes are in demand and which demographic is buying them. A predictive machine learning model and a visualization analysis of customer demographic would provide the sufficient insight in what customers are buying what size.

## Customer Summary

The purpose of this application is to provide a deeper analysis and understanding of the customer base. The intended customer that we will develop the application for are the manufactors, stakeholders, and directors of Bob’s Bike Shop. Right now, they are having trouble understanding what jersey size shirts the customers want and they have come to us for help. These analysis tools are intended to provide insight on what jersey sizes are in the most demand, and which customer demographic wants them. By offering these descriptive and perscriptive analysis tools, like bar charts and non-descriptive machine algorithms, to our directors, they can make better decisions on which jersey size to prioritize. This will lead to a better budge allocation efficiency and less resources wasted.

## Existing System Analysis

The client currently uses a SQL Server relational database to store their sales records. They may have the records, but they still need visualization tools to gain a better understanding of what these records entail. We would need to use the database to feed the non-descriptive machine learning algorithm. We can also create visualization graphs like bar graphs to find out the main demographic of Bob’s Bike Shop’s customers. This would make it easier to read and interpret the data, therfore making better decisions.

## Data

We will be using the dataset ‘Bike Sales in Europe’ from kaggle.com as a test to ensure that the machine learning algorithm is working correctly. In order to filter out the necessary data regarding classic jersey sizes, we will be using SQL Server Management Studio. We will query only the neccessary data regarding classic jerseys. This will leave out any other products such as bikes and accessories and allow us to only focus on the specific jersey sizes, the main problem at hand. As more records will be added in the future, we can continue to update the csv table set from the database and therfore updating the machine learning algorithm to better fit the demographic.

## Project Methodology

In this project, we will be using the waterfall methodology to complete this project. Even though this method may be more rigid compared to other methods, it clearly deines the requirements needed to complete this project. The scope and scale of the machine learning algorithm will match well with this method. In the waterfall methodolgy method, we will include the requirements phase, the design phase, the implementation phase, the verification phase, and the maintenance phase.

In the requirements phase, we will set up a meeting with database administrators and directors to confirm and collect the necessary data used to create the machine learning algorithm. We will also gather other datasets such as the ‘Bike Set in Europe’ from kaggle.com in order to train the machine learning model. If the data needs to be cleaned up, we will use SQL Server Management Studio to properly query the necessary data. By using the water methodolgy method, we can clearly define the requirements and necessary tools to avoid scoop creep.

In the design phase, we will begin to develop a user friendly interface for the application. We will also begin finding the correct classification method for the machine learning model. As of right now, the plan is to implement a decision tree model. The finished application will most likely be in the form of a Jupyter notebook.

In the implementation phase, we will begin implementing the machine learning model in python. The python version used is 3.11 on Jupyter Notebook 6.4.12. The operating system will be windows 10. The modules used within the python program will be pandas 1.4.4, scikit-learn 1.0.2, numpy 1.21.5, and matplotlib 3.5.2. Each of these modules will be used to import the data, implement the machine learning model, and visualize the data.

In the verification phase, we will analyze the data and accuracy of the model. This will be done using the matplotlib module. We will use bar graphs to show a visualization of the data and we will use the confusion matrix to analyze the accuracy of the machine learning model. From these insights, our directors will be able to make better decisions.

In the maintenance phase, we will continue to improve the model by adding in more data as time progresses and continue to monitor the application for any bugs and issues.

## Project Outcomes

The final outcome for this project will be done in a Jupyter notebook. The application will provide a section for the customer to input the variables (i.e age and gender) and the machine learning algorithm will predict the classic jersey shirt. In addition, the application will provide bar charts describing the data. Finally, the customer will be able to run an accuracy test of the model as well as view a confusion matrix describing how well the predictive machine learning model did. If the customer wishes to view the SQL queries used to filter the data, those files will also be provided.

## Implementation Plan

The implentation plan will first involve querying the correct data. By filtering out the unecessary data, we can speed up the process of the training model and reduce indesirable outcomes. We will use SQL Server Management Studio 19 to query the data. After the data has been queried, we will save it in a CSV file named ‘Sales Jerseys.csv’. From there, we will create a Jupyter notebook using Jupyter Notebook 6.4.12 and code the machine learning model in Python 3.11. The module that we will use to read the ‘Sales Jersey.csv’ will be pandas 1.4.4. We will then use scikit-learn 1.0.2 to implement the decision tree classifier and train the machine learning model useing the csv file. The benefits of using the decision tree classifier is that it is easy to implement and does not require normalization or scaling of data. It is also an intuitive model for stakeholders. Lastly, we will not be required to predict continious data, this classification model will be optimal for this project. After the model has been implemented, We will visualize the data using matplotlib 3.5.2. The plan is the use two bar graphs for genders and ages separately, and those will determine the dominant demographic for each jersey size. Finally, the machine learning algorithm will run through an accuracy test and a confusion matrix will be added to describe how well the tests ran.

The phases of the rollout and bug testing will be performed through a thorough code inspection by a group of 3 or 4 people. These people will go through rigorious testing to find any bugs. After all the bugs have been fixed, the software will go through the directors and they will be able to use the application effectively.

The dependencies of this project includes used in this project is described above. In a nutshell, the main dependencies that we will be using will be pandas 1.4.4 for importing data, sci-kit learn 1.0.2 to implement the machine learning model, and matplotlib 3.5.2 to visualize insights. The project will be done in Jupyter Notebook 6.4.12 with Python 3.11. The operating system that will be used to complete this project is Windows 10.

## Evaluation Plan

The verification methods are essential in the development of an appliction. In order to ensure that the entire application runs properly, unit testing will be done in each development of the stage. The unit testing will include things such as importing and implementing the machine learning model, adding each bar graph describing the data, adding an accuracy test of the model, and adding the confusion matrix describing accuracy of the machine learning model.

The validation strategy used to determine the accuracy of the model will involve splitting the data between a 80% and 20% ratio. The 80% data will be used to train another model and the 20% data will be used to test the model. If the accuracy results show a result of around 0.5 or more, that would be a success.

## Resources and Costs

Since the dataset used in this project is free and open-sourced, there will be no cost of charge. The software required to complete this project, like the Jupyter Notebook and SQL Server Management Studio, will also be free. The costs of the application development will be $200 per hour. Since the estimated project time will be around 40 hours, the charge for application development will be $8,000. If you also account for the hardware and operating system required to complete this project, that will add another $2,000 to the project. Finally, an aditional $2,000 will be used for debugging purposes, re-iteration cycles, and deployment. This will lead to a total cost of $12,000. If the application must be maintained and updated every year, that would be another $2,000 for application maintainability. The total cost of the entire project will be $12,000 with $2,000 per year for maintainability.

## Timeline and Milestones

The project will start on June 1st and the estimated completed date will be June 17th. The estimated time to complete this project is 40 hours.

|  |  |  |  |
| --- | --- | --- | --- |
| Milestone | Start – End Dates | Duration | Resources Assigned |
| Requirement Analysis | June 1st – June 2nd | 5 hours | Directors, Manufacturers, Customers, Stakeholders |
| Design of application | June 5th – Jun 7th | 5 hours | Software Engineer |
| Implementation and code | Jun 7th – June 12th | 15 hours | Software Engineer |
| Verification phase and debugging | June 13th – June 16th | 10 hours | Software Engineer, Quality Assurance |
| Project Delivery | June 17th | 5 hours | Software Engineer |

# Part C: Application

Part C is included in the submitted files.

A screenshot of a computer

Description automatically generated with medium confidence

# Part D: Post-implementation Report

## A Business (or Organization) Vision

The purpose of this project was to figure out which jersey sizes are in demand by customers. In the application, you are given a non-descriptive machine learning model that predicts the most likely jersey size a customer would buy given their age and their gender. You are also given several visualization tools that describe the customer demographic and accuracy of the model. A stakeholder or a director can use these tools to better allocate resources that would increase the supply for in demand jersey sizes.

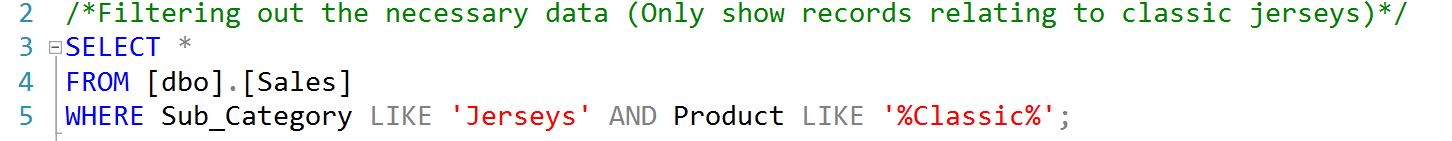
Let’s say you wanted to target a demographic of 34 year old males. You would simply input those characteristics into the machine learning model, and it would predict the most likely jersey size shirt they would buy. From there, directors and stakeholders could direct their marketing strategy towards that demographic and produce more of those jersey sizes.

## Datasets

The dataset used in this project was the ‘Bike Sales in Europe’ dataset from kaggle.com. (<https://www.kaggle.com/datasets/sadiqshah/bike-sales-in-europe>). This dataset contained the sales purchase of all products, including bikes, waterbottles, and other accessories.

A picture containing text, number, parallel, document

Description automatically generatedHowever for the purpose of this project, we only needed the sale records regarding classic jersey sizes. In order to filter out the other sales, I used SQL Server Management Studio 19 to query the data regarding classic jersey sizes.



A picture containing text, screenshot, parallel

Description automatically generated

I then saved the results of the query named ‘Sales Jerseys.csv’. This csv file was used to train the final application. The original file can be found under Part D, named ‘Sales.csv’. The orignal file was NOT used in the final application.

## Data Product Code

As mentioned above, processing the raw data involved querying the records regarding classic jersey sizes. In order to get the gender and age demographic of the customer base, I used the same querying techinque to count up all the records and group them based on gender and age respectively. I then added those results to bar graphs to provide a better visualization of the data. One bar graph described the number of males and females who bought each Jersey size, and the other bar graph described the age demographic for each jersey size. These were known as the descriptive methods. (NOTE: The queries used in this project are saved under “SQLQuery1” in the project files.

A screenshot of a computer

Description automatically generated with low confidence

A screenshot of a computer screen

Description automatically generated with low confidence

For the non-descriptive method/predictive method, I used a decision tree classifier to predict the jersey shirt size. This could also be known as a supervised machine learning model. There were several benefits by using this techinique. First, the algorithm is relatively simple and easy to understand. This would make it easier to explain to stakeholders and provide high maintainability. Also since the predictions were going to be non-continous, there was no reason to implement a continious prediction models such as lienar regression. The decision tree classifier was effective for the purposes of this aplication. As for the data, it was trained using the Sales Jersey.csv file, NOT the Sales.csv file. The X values included in this project was the ‘age’ and ‘gender’. The Y values included in thie project was the ‘Jersey size’, or products column.

A screenshot of a computer

Description automatically generated with medium confidence

In order to determine the accuracy of the model, I split the dataset into two groups: The training set and the testing set. The ‘X\_train’ and ‘Y\_train’ would be used to classify and train the model, while the predictions for the ‘X\_test’ variable would be compared to the ‘Y\_test ’ variable. The more predictions that match the ‘Y\_test’ variable, the more accurate the model was.

## Objective (or Hypothesis) Verification

The objective of this project was to create a machine learning model that predicts which jersey size a customer would buy based on their age and their gender. The accuracy goal for this project was around 75%. Unfortunately, the objective fell a little short with an accuracy goal of around 50%. This may be caused by a lack of records needed to train the data. Another theory is that the gender and age of the demographic may not have as a strong correlation to what size they will buy.

## Effective Visualization and Reporting

These visualization tools would allow stakeholders and directors to produce more of a certain size to target certain demographics. Foor example, the first bar graph describes the number of males and females. If we wanted to target the male demographic, we would most likely produce XL classic jersey sizes and S jersey sizes. If we wanted to target the female demographic, we would most likely produce S jersey sizes and L jersey sizes. The same techinque can be applied to the second bar graph. If we wanted to target adults, we would most likely produce S jersey sizes or L jersey sizes. If we wanted to target the youth group, we would most likely produce small are XL jersey sizes.

A picture containing text, screenshot, diagram, font

Description automatically generated

A picture containing text, screenshot, diagram, plot

Description automatically generated

The last visualization tool was a confusion matrix used to describe the accuracy of the supervised machine learning model. Even though the accuracy objectives were not met, these visualization tools have effectively described what the main target audience is. Most of the customers buying jersey sizes are male adults. We could use this insight to produce more jersey sizes for that demand. And as more sales records regarding jersey sizes are added, there will be more data for the machine learning model to train on. This would lead to an increase in accuracy of the model. The confusion matrix would describe the accuracy more and more.

A picture containing text, screenshot, diagram, rectangle

Description automatically generated

## Accuracy Analysis

In order to measure the accuracy of the model, we use the Scikit-learn module function ‘test\_train\_split’ to split the data into training sets and data sets. As mentioned in the ‘Objective (or Hypothesis) Verification method’, The ‘X\_train’ and ‘Y\_train’ would be used to classify and train the model, while the predictions for the ‘X\_test’ variable would be compared to the ‘Y\_test ’ variable. The more predictions that match the ‘Y\_test’ variable, the more accurate the model was.

A screenshot of a computer program

Description automatically generated with low confidence

According to the Scikit funciton, the accuracy of the model is around 52.689%. This does not match the goal of 75%, but that will most likely change as more sales records are added into the future. When there is more data to train on, there is better accuracy from the machine learning model.

## Application Testing

Since Jupyter Notebook allowed for certain sections of code to be executed, This would allow for unit testing after every major section in this project. In this project, I separated the code between building the machine learning model, adding in the bar chart for gender, adding in the bar chart for age, adding in the accuracy test, and finally adding in the confusion matrix. I made sure each section was operating effectively before moving on to the next section. After all the units were working properly, I went back to the beginning and began a full walkthrough of the application. If there were any bugs or visualization errors, those would be updated and tested again. Since this project was relatively small, it was easy to implement tests and make changes quickly.

## Application Files

The ‘app.ipynb’ contains the main application and machine learning model. The ‘C964 task 2.docx’ is the document for this project. The ‘Sales Jerseys.csv’ file is the cleaned up data that the applciation uses. The ‘Sales.csv’ file is the raw, unfiltered data downloaded from kaggle.com. Finally, the ‘SQLQuery1.sql’ is the file used to query and gather the data to train the model and make the visualizations.

A screenshot of a computer

Description automatically generated with medium confidence

## User Guide

1. Make sure these programs are installed before running the application
   1. Anaconda 3
   2. Jupyter Notebook 6.4.12
   3. Python 3.11
   4. Sci-kit learn (python module)\*
   5. Numpy (python module)\*
   6. Pandas (python module)\*
   7. Matplotlib (python module)\*

(NOTE: All the python modules should be installed automatically if Anaconda 3 is installed. If that is not the case, you may need to manually install them through windows command prompt)

1. Make sure you have the files downloaded from the WGU assignment
2. Open up Anaconda 3 and launch Jupyter

A screenshot of a computer

Description automatically generated

1. Locate the file you downloaded and launch ‘app.ipynb’

A screenshot of a computer

Description automatically generated with medium confidence

1. After you launch the application, you can add in the variables and the machine learning model would predict the likely jersey size they will buy. The **first paramenter input is the age** of the customer, and the **second parameter input is the gender** of the customer.
   1. If you want to predict a **34 year old male**, you would type ‘model.predict([[**34,0**]])’ under user input.
   2. If you want to predict a **70 year old female**, you would type ‘model.predict([[**70,1**]])’ under user input
   3. Press ‘**Shift + Enter**’ to show the output.

A screenshot of a computer program

Description automatically generated with low confidence

1. You can view the bar chart describing the customer demographic by pressing ‘**Shift + Enter**’ under the 2nd and 3rd code block.

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

1. To test the accuracy of the machine learning model, you can execute (Shift + Enter) the ‘Accuracy test’ block code, and it will display how accurate the model is. (The closer to 1, the more accurate the model is)

A screenshot of a computer program

Description automatically generated with low confidence

1. You can also view a confusion matrix describing the accuracy test that was just executed. (The last code block)

A screenshot of a computer program

Description automatically generated with low confidence

## Summation of Learning Experience

Machine learning and data analytics has been a topic I have been studying for several months now. My academic experience with SQL Server has allowed me to filter out the necessary data from kaggle.com. I have also used this tutorial <https://youtu.be/7eh4d6sabA0?t=2> to help download the softwares and modules needed for this project.

I would also say that the ‘Data Strucutres and Algorithms 2’ and the ‘Software Quality Assurance’ classes helped me tremendously in this project. ‘Data Structures and Algorithms 2’ has solidifed my skill in coding comfortably in python, while ‘Software Quality Assurance’ has taught me the proper terminology used to write this document.

By working on this project, I have learned many lessons. I was able to apply my data analysis skills for SQL as well as implementing my communication skills. Trying to communicate in a non-technical and technical way is a cruical skill in the workplace, and I am glad I was able to exercise this skill. I also learned how to utilize outside resources such as kaggle, google, and youtube. I will use these new found skills and apply them in my career as a data analyst in the future.

Sources

Hamedani, M. (2020, September 17). *Python machine learning tutorial (data science)*. YouTube. https://youtu.be/7eh4d6sabA0?t=2

Shah, S. (2020, September 23). *Bike sales in Europe*. Kaggle. https://www.kaggle.com/datasets/sadiqshah/bike-sales-in-europe