|  |
| --- |
| MS in Applied Data Science  Project Portfolio |
| The 5 Biggest Data Science Trends In 2022 |
| March 12th, 2022  Syracuse University  IST 782  Noah Laraway |

Table of Contents

* Report Overview (page 1)
* Portfolio Project Descriptions (page 2)
* Learning Objectives (page 4)
  + Describe a broad overview of the major practice areas of data science (page 4)
  + Collect and organize data (page 5)
  + Identify patterns in data via visualization, statistical analysis, and data mining (page 8)
  + Develop a plan of action to implement the business decisions derived from the analysis (page 11)
  + Develop alternative strategies based on the data (page 13)
  + Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization (page 15)
  + Synthesize the ethical dimensions of data science practice (page 16)

Report Overview

This report supports the project portfolio milestone requirement for the M.S. in Applied Data Science at Syracuse University. It provides details and reflects on projects completed and how they tie together with the overall learning objectives of the program. Link to supporting files:

<https://github.com/larawayn/Portfolio_Milestone>.

Portfolio Project Descriptions

IST 707 – Applied Machine Learning

This project was a fashion image recognition project that takes photo data for clothing and uses machine learning and image recognition to predict the class label for that image, e.g., dress, sneaker or coat. Python was utilized to perform nine different machine learning models and a recommendation of the best model was made based off an analysis of the model scores.

IST 659 – Data Administration Concepts and Database Management

This project consisted of creating a golf score database from conceptual design all the way to production using SQL. Three years of prior personal golf score data was used including players, courses, dates and player scores. The data was organized into diagram tables to show how it was related and what the overall database should look like. SQL was then used to create the database and an analysis of the data was performed to answer questions about how golf scores changed over time, by course and by what players were in the golf outing.

IST 687 – Introduction to Data Science

This project was an analysis of craft breweries, beer styles and their specifications in the US. Brewery and beer data was gathered from multiple sources and loaded into R. The data was then cleaned, transformed and further prepared for analysis. The analysis of the data utilized aggregate functions and modeling to produce statistics and visualizations to make conclusions about the data. The results showed which beer styles are the most popular, which states and cities brew the most beer, which states produce beer with the highest alcohol content and bitterness and modeling results that described how alcohol content and bitterness are related.

MBC 638 – Data Analysis and Decision Making

The project for this course was a process improvement plan for reducing household expenses. The main objective was to gather data and draw insight on a previous process and then improve the process through statistical analysis. Twelve months of prior spending data was gathered and analyzed. After the initial analysis, the improvement plan was implemented with a goal of reducing monthly spending by $200. The improved process was monitored and compared to the original process to statistically confirm that the process was had indeed improved and the goal was met.

IST 652 – Scripting for Data Analytics

This project was a real estate analysis project that looked at 20 years of home sale prices for Denver, CO and the US as a whole. The main focus was on pulling large amounts of data from multiple sources and using Python to analyze the data, create visuals and answer technical questions about home sales trends. The results described how real estate prices in the US and Denver have changed overtime and also went into detail on the top and bottom Denver neighborhoods for median sales price over the timeframe.

IST 623 – Introduction to Information Security

This project was on the cyber security threat of killware. Killware is a threat where hackers aren’t looking for monetary gain, but are trying to cause harm or even death to people. The main focus of the project was on the background of killware, what industries are at risk, the technology used and how security breaches occur. An in-depth investigation was also performed looking at two recent killware attempts.

Learning Objectives

Describe a broad overview of the major practice areas of data science.

There’s been a growing trend of companies in all industries attempting to utilize their data to add value to their business. Through my project in the Applied Machine Learning course, I showed how data science and machine learning are used for image recognition. In the fashion industry in particular, applications for machine learning can be to search and tag clothing and also for deciding the correct measurement of shoes and attire. This has led to the adoption of image recognition in the fashion industry. To start the project, the image data was brought into Python to analyze.

Machine learning fashion image recognition example images

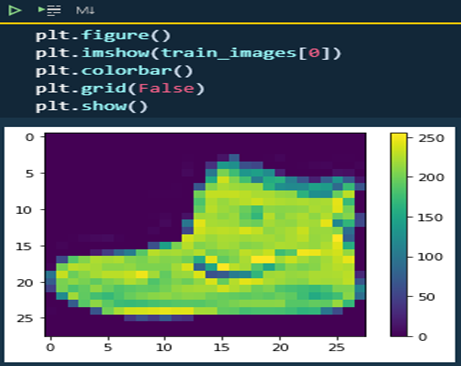
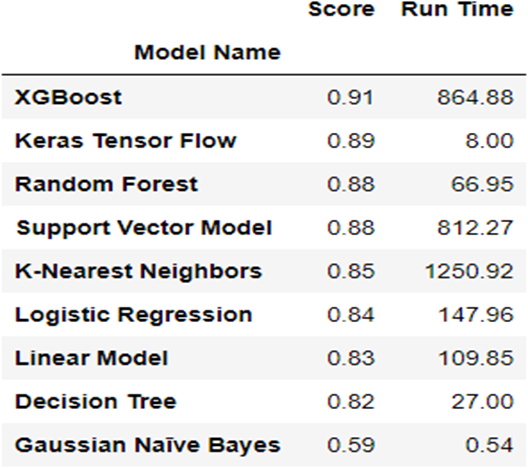


Image data was ran through nine machine learning algorithms to compare the results and decide on which algorithm is best suited to be utilized.

Model score and run time table sorted by score 

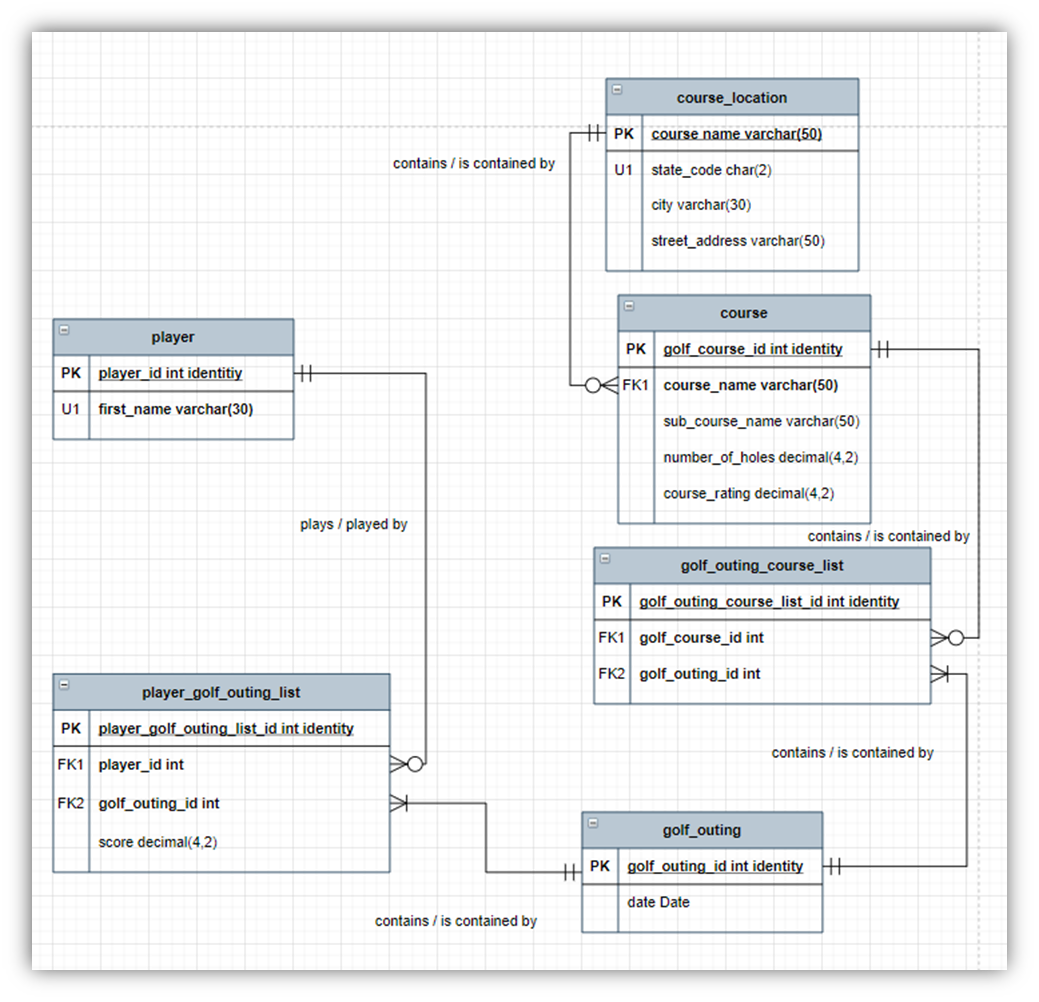
The XGBoost model offered the best score at a 91% prediction level. This project shows that by using machine learning for image recognition we can minimize the time needed to process and categorize images.

Collect and organize data

In data science it is essential to have a good understanding of the data you’re working with before doing any analysis. Knowing the structure, data types and how different data sets are related is vital to avoid potential issues when getting further along in the project. Through my project in the Data Administration Concepts and Database Management course, I demonstrated how to collect data, relate tables and create a database that was organized in a way that would be simple to pull and manipulate information. For this project I created a golf score database to help me keep track and ultimately find out how much my golf game has improved over time.

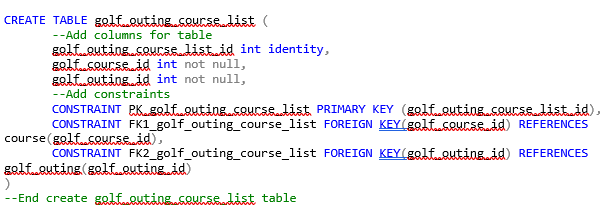
The project started with collecting golf score data from various sources including old score cards, data kept on a phone app and also pulling information from golf course websites. At that point I had to decide what information I wanted to get out of the database as a user to understand what data needed to be included in the database. I wanted to know how my golf score changed over time and how different variables make an impact on my score such as the course, time of year the round was played, who I played with and how many holes were in the outing. To do this I needed information about the course, players, scores and the golf outing. A conceptual model was created to organize the data and put the correct columns into tables. After that a logical model was made to add primary and foreign keys and show the table relationships. Bridge tables were also added to eliminate any many to many relationships.

Logical model diagram for the Golf Score Database project



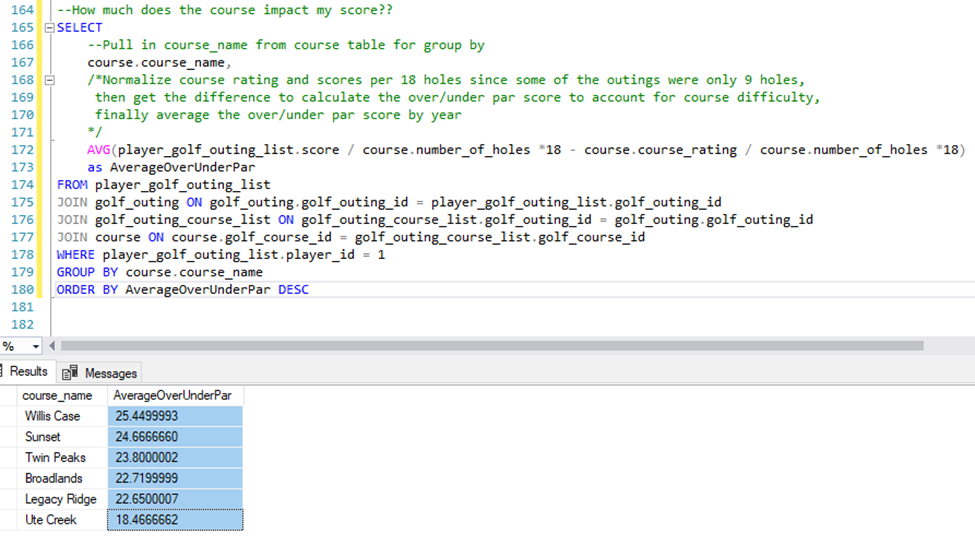
SQL was then used to create the tables, apply the table constraints and insert the data.

Example of table created in SQL for Golf Score Database project.



Once the database was created, views and select statements were used to analyze the data.

Example of a select statement in SQL to show how scores changed depending on the golf course



This project helped me understand how important it is to do additional work upfront when collecting and organizing data to prevent issues during the analysis. I initially had certain columns with the wrong data type that needed to be adjusted so that aggregate functions could be performed correctly. Knowing the data, what format it needs to be in and how the tables are related is very important when creating a database.

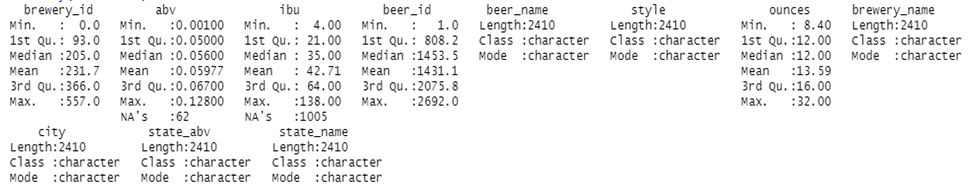
Identify patterns in data via visualization, statistical analysis, and data mining

A large focus throughout the courses in the program has been to identify patterns with data. Being able to draw insight from data patterns is important during an analysis to help draw conclusions. My project in the Introduction to Data Science course showed my understanding of using visualization, statistical analysis and data mining to identify patterns. The project for this course was a craft beer analysis and focused on analyzing beer and brewery information to produce statistics and visualizations to make conclusions about the data. The datasets contained information on the brewery name, brewery location, beer name, beer style, alcohol content (abv) and bitterness (ibu) of the beer.

The main analysis was finding out which beer styles were the most popular, which cities and states had the highest number of breweries and beers and what cities and states produced beers with the highest bitterness and alcohol content. Another aspect of the project was to model how alcohol content is related to bitterness.

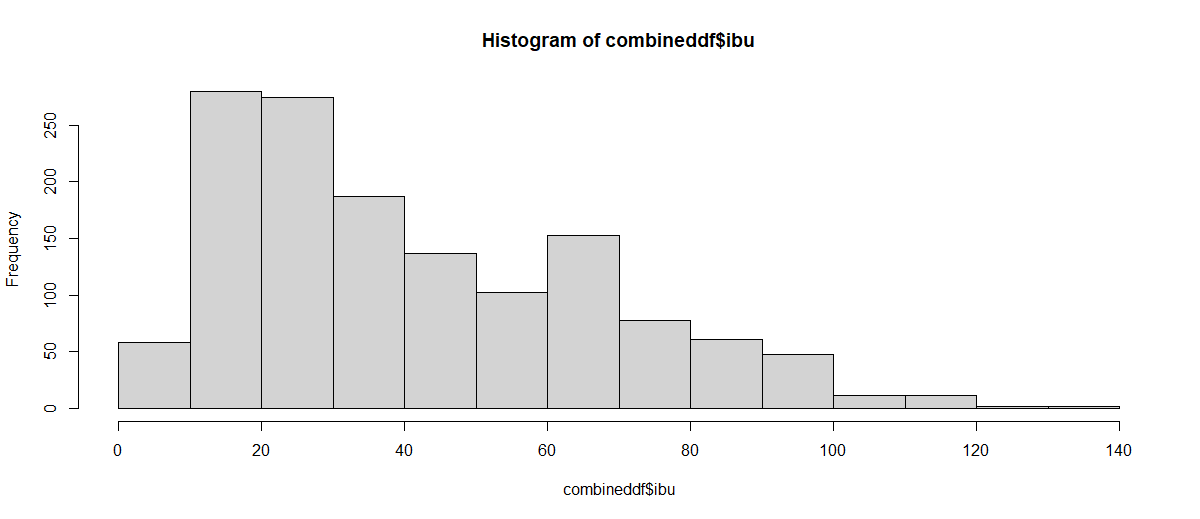
When starting any data project, it is important to get a statistical understanding of the data before beginning an analysis. For this project summary statistics and histograms of the data were created.

Summary statistics for Craft Beer Analysis Project



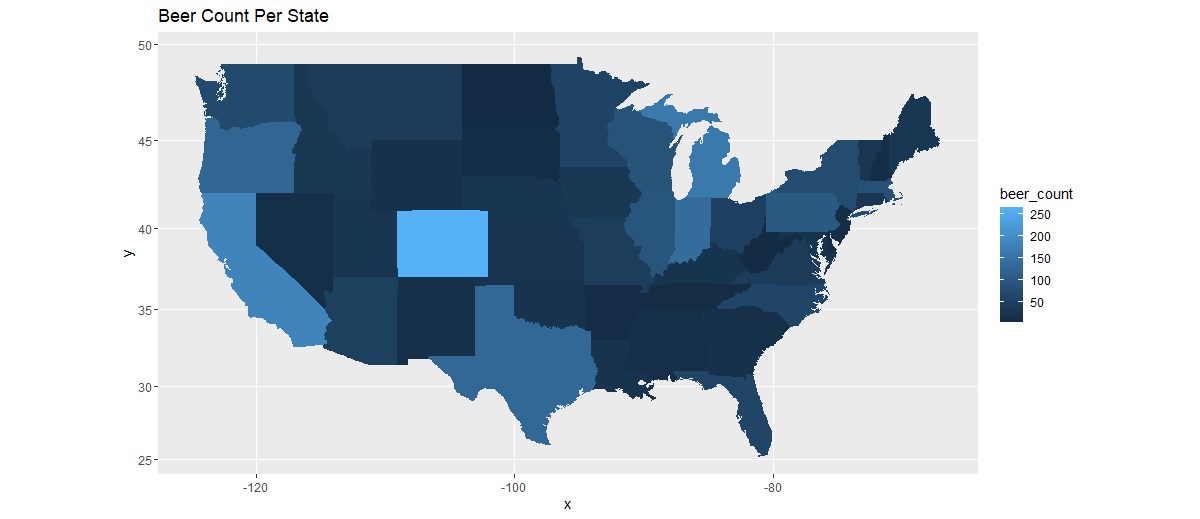
Looking at the mean, median and range of the descriptive statistics helps draw some initial insight on what to expect during the analysis. From the statistics summary the median alcohol by volume is 5.6% and the median IBU is 35 which are fairly typical for craft beers.

Histogram of bitterness content (IBU)



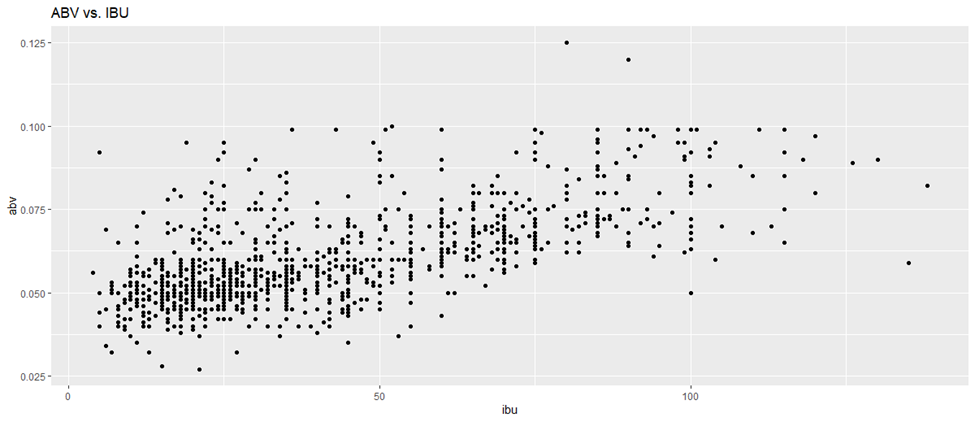
For IBU the highest frequency is between 10 and 30 IBUs, but the histogram is skewed towards the right meaning there are some beers in the dataset with relatively high IBU content.

To find out which states in our dataset produce the most beer a map was created to visualize the data.



From the visualization, Colorado is the state with the highest beer count.

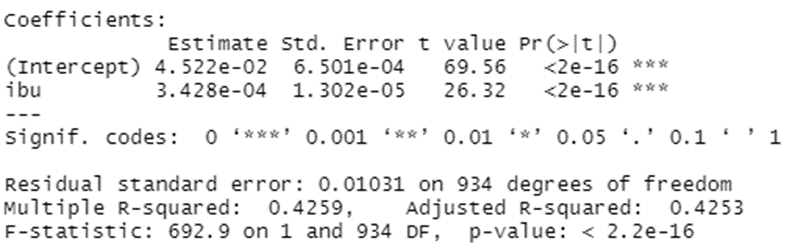
To visualize how alcohol and bitterness content are related a scatter plot was created.



This plot shows a general trend with an increase alcohol content and increase bitterness.

A linear regression model was used as well to show the relationship between the two variables.

Summary of linear regression



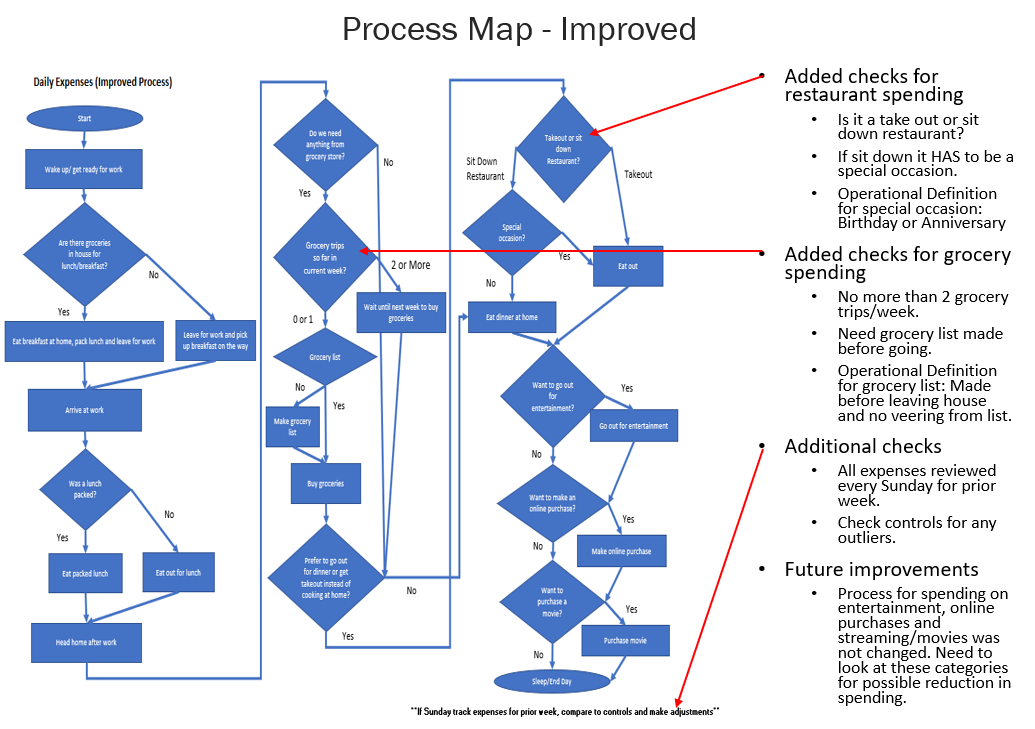
The linear regression summary shows an adjusted R-squared of 0.4253 with a p-value of 2.2e-16. This R-squared means that 42.53% of the variation in ABV can be explained by variation in IBU. Also, the p-value is very low (less than 0.05) so this equation is significant.

Looking at the summary statistics, creating visualizations and performing modeling techniques is very important when trying to understand how the data is related. Using multiple techniques is often needed when looking for patterns to draw conclusions about the data.

Develop a plan of action to implement the business decisions derived from the analysis

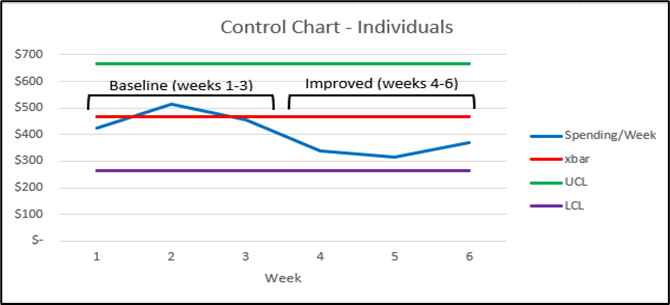
When doing a data analysis project, it is important to understand what the ultimate goal is and how the conclusion and insight can be used. It is also important to make sure any plan of action from the data analysis is followed through and monitored. In the Data Analysis and Decision-Making course, I designed a process improvement plan that focused on reducing the household expenses for my family. I gathered spending data for one year and broke the data into essential and discretionary spending. After analyzing the discretionary spending, I discovered that too much money was being spent on restaurants and groceries. A process map was created for the improvement plan that added checks before a restaurant or grocery store purchase was made.

Process improvement map



The process map contained added questions to be asked before making purchases to limit spending such as “Has a grocery list been made?” or “Is the restaurant trip for a special occasion?”. Based on the answers to these questions an additional requirement had to be satisfied or the purchase was not made.

Control charts were created as well to monitor the new improvement plan.



The control charts showed an improvement in reducing spending compared to the prior baseline spending habits.

When performing data analysis and drawing conclusions it is crucial to make sure the plan of action is implemented properly. There needs to be a clear process in place that is monitored so that the recommended business decision is implemented properly.

Develop alternative strategies based on the data

When working on projects, I’ve found that often times issues with the data will arise that require workarounds to complete the project. If there are concerns that the data doesn’t look right, additional investigation and decisions need to be made on how to move forward. Data issues were discovered while doing my Real Estate Analysis project in the Scripting for Data Analytics course.

This project focused on real estate prices for Denver, CO and the US as a whole. One focus of the project was finding how prices for neighborhoods in Denver had changed overtime. When looking at these neighborhoods it was discovered that one of them had a huge drop in median sales price from 2010 to 2021.

Data frame for bottom 10 neighborhoods by dollar change:



The data frame shows that there is an issue with the neighborhood N Green Valley. It says the median sales price in 2010 was over $17MM, but in 2021 it was only $426K.

After further investigation I discovered that this neighborhood is relatively new and was in the early stages of development in 2010. Most likely the 2010 sales were due to home builders either buying multiple lots in one sale or buying large amounts of acreage in one sale. I decided that the data for this neighborhood had to be removed.

When dealing with large data sets there can often times be issues with the data. Being able to know what to look for and finding anomalies is important to understanding the data. When making an accurate analysis, further investigation on suspect data and making decisions on what to do with that data is often needed.

Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization

Being able to communicate findings from a data analysis project is one of the core concepts in the field of data science. If you aren’t able to effectively report and present recommendations and conclusions, the hard work of data collection, mining and analysis can end up meaningless. It’s important to be able to construct your finding in a concise and clear manner that includes project background, visualizations and interpretation.

Throughout the program I have performed numerous class presentations. At the end of the Applied Machine Learning, Scripting for Data Analytics and Introduction to Information Security courses I presented the project that was worked on in the course. These presentations were beneficial in understanding how to get information about my project and the analysis communicated effectively.

The presentation for the Machine Learning course started with a background on image recognition and how it relates to the fashion industry. An overview of the data was then presented followed by results from the machine learning models. The presentation concluded with a recommendation to use the XGBoost model for image recognition on the data set that was used.

The presentation for the Data Scripting for Data Analytics course on Real Estate Analysis began with an overview of real estate trends and how prices have increased overall recently. The data sources and a description of the program were then discussed in detail. The data analysis was presented next covering price changes for the US and Denver. The final results and conclusion discussed how real estate prices have increased drastically across the US and particularly in Denver.

The Information Security presentation focused on how computer hackers used killware. A brief introduction on killware and how it is used to cause harm was explained. An overview of the industries at risks were discussed followed by an in depth look at two recent killware attempts and the technologies used by the hackers. A final conclusion was then presented covering the future of killware and what organizations can do to help prevent it from happening.

Overall, these presentations showcased the projects from the courses and the analysis that was performed. On completion of the presentations, I gained a better understanding of how to structure information, data and reports to get my point across to the audience. Making sure the audience understands the background of the project, the data analysis and the overall conclusion or recommendation is crucial in effectively communicating the results of a project.

Synthesize the ethical dimensions of data science practice

An important aspect of data science ethics is keeping data and systems secure to prevent a security breach. Security breaches can leak sensitive information about individuals or companies that can lead to identity theft, technology being stolen or ransomware attacks. Some security breaches can even lead to deadly consequences. My project for the Introduction to Information Security course, looked at security breaches in the form of killware which is where hackers are trying to cause harm or even death.

The project focused on two recent events and covered the background of killware, technologies used and how the security breaches occurred. One of the recent events was at the Oldsmar water facility in Florida where a hacker broke into the system and attempted to adjust a concentration of one of the chemical additives to a level that would make the water toxic to humans. The breach occurred due to a new software that had been installed on the plant’s computers to allow employees access when they were working remotely. Fortunately, an employee was watching the computer monitor when the mouse move on its own in an attempt to alter chemical concentrations. The employee notified management and the killware attempt was stopped with no harm done.

The other killware event was at a petrochemical plant in Saudi Arabia. The hacker broke into the system and started to make the equipment at the plant operate abnormally. Upon investigation it was discovered that the goal of the attack was to cause an explosion at the plant. Thankfully the code that the hackers used had an error in it and the attack did not go through as planned.

This course and project helped me to understand the importance of security to prevent a breach. As technology continues to change it is vital to make sure that computer systems are secure and not compromised even if it adds cost to the company. Hackers are constantly improving their techniques for gaining security access and having the most up to date software and security features should be a top focus for companies going forward.