           The Master of Science in Applied Data Science Program at Syracuse University aims to create Data Science Professionals capable of demonstrating several skills gained through courses designed around the Program's Learning Goals. As a student participating in the program, the "applied" aspect was explicitly appealing; course projects allow students to actively demonstrate their growth as Data Scientists. The Program's Learning Goals keep students on track to becoming successful Data Scientists.

           Working with data requires two crucial components to set up data users for accurate and successful analysis and decision-making: storage and easy access. The Data Administration Concepts and Database Management course highlighted the importance of designing a well-thought-out database; the design process involves various steps, including dissecting the business processes and end-user interface. Online diagram software like <https://app.diagrams.net/> allowed for document preparation, such as producing Entity Relationship Diagrams. SQL coding, including Data Definition Language (DDL), Data Manipulation Language (DML), and Data Control Language (DCL), generated the database and allowed for data retrieval.

As a course requirement, students produce a project including diagrams, a coded database, data questions, and a user interface. A friend's real estate company needed a better tracking system for sales transactions. Ashley Lew started Property by Ashley to represent sellers and buyers in California home sales transactions. At the time, Ashley used her website to track transactions. The data was collected from the website to create a database for her company, Property by Ashley. The goals of the Property by Ashley database are[[1]](#footnote-1):

* Provide a structured, efficient, and effective way to track sales transactions.
* Provide Ashley with valuable information about her business to help her plan for future transactions.

When designing the Conceptual and Logical models, it was crucial to think like a Real Estate Agent: What questions are Real Estate Agents interested in when assisting buyers and sellers with home sales transactions? It was also essential to consider the following: How should the data get presented to a Real Estate Agent? Ashley Lew is not a SQL user. Therefore, installing SQL Server Management Studio on her laptop and giving her a list of SQL queries is an illogical option. The project emphasized the value of considering the business rules and end users while designing a database. Additionally, the project highlighted the significance of meaningful questions to answer with the data collected from her website; Ashley was instantly excited to hear the data questions because she saw how the answers would add value to her business.

Producing a normalized model was particularly challenging. The concept of normalization was foreign but realized when it was time to write SQL code to create the Property by Ashley database. The data redundancies became evident, so normalization changes occurred. The task emphasized preparing and structuring data for storage in a way that conceptually makes sense and test queries for efficient data retrieval. Furthermore, the Logical Model is a live document that changes as new business processes and other factors become a consideration.

Data Visualizations are essential tools for Exploratory Analysis. Data Visualizations help reveal meaningful relationships and patterns between variables in data without building Machine Learning and Statistical Models. There is an art involved in producing Data Visualizations. There are rules to avoid breaking, like selecting the wrong visualization to answer a specific question and information overload. In general, Data Visualizations provide an opportunity to tap into the creative side and present information as a work of art.

The Information Visualization class heightened abilities to produce simple, meaningful visualizations that invite questions and inspire curiosity. Students must analyze a dataset and produce a poster with various visualizations to tell a story. A California Wildfire dataset, hosted by Kaggle and previously collected through CAL Fire's website, was the basis of the story. The poster[[2]](#footnote-2) aimed to inform California newcomers of potential wildfire danger zones by highlighting Counties based on historical incidents. The analysis used the programming language R to generate all visualizations. The main visualization is a map of California denoting acres burned during wildfire incidents. Supporting visualizations included bar graphs of Structures Damaged and Destroyed by wildfire incidents, a word cloud describing wildfire locations, and a trend analysis of acres burned yearly.

Creating a layout for the poster using Adobe Illustrator was explicitly challenging. When learning Residential Architectural Lighting Design a few years ago, AutoCAD was the tool selected to create lighting drawings for high-end homes. Adobe Illustrator is similar to AutoCAD; however, creating the poster felt more like designing the house than just laying out the lighting fixtures. Choosing the right amount of information to showcase was also tricky. Nevertheless, creating and presenting the poster was a fun experience; it allowed for tapping into the creative side and using visualizations to tell the story behind a dataset. Most importantly, it was necessary to consider the audience and ask, "What is the easiest way to communicate a powerful message?"

As a student of the Applied Data Science program, one of the learning goals encourages students to create actionable insights across societal, business, or political contexts using data and the data science life cycle. The Data Analytics/Applied Machine Learning course presented the opportunity to explore societal issues through its course project. There is currently an opioid addiction problem across the United States, and it turns out the problem stems from doctors overprescribing opioids to patients. Unfortunately, some doctors purposefully overprescribe opioids because of greed. Our team utilized data analytics[[3]](#footnote-3) to identify anomalies categorized by high opioid prescriptions unusual for a specific prescriber type, potential high-risk States, and prescriber types in the Opioid Prescriber dataset provided by the Center for Medicare & Medicaid Services.

**Capture:** Kaggle is hosting a subset of the Opioid Prescriber dataset. The team decided to obtain the most recent data by querying the online database provided by the Center for Medicare and Medicaid Services. After executing a query for the most recent data, the online database created a CSV file.

**Maintain:**Kaggle provided an R-script for data preprocessing steps. The team modified the script and used it to perform any cleansing steps. The label generated for each prescriber based on the drugs prescribed is opioid prescriber = True or opioid prescriber = False. The transformed data frame was exported to a CSV file and used to build all models.

**Process:**The team utilized Clustering and Classification methods, such as Naïve Bayes, to identify States and Prescriber Types with high conditional probability values for opioid prescriber = True and Decision Trees to detect drugs contributing to the opioid prescriber = True label.

**Analyze:**Exploratory/Confirmatory methods like Association Rule Mining helped confirm high-risk States and prescriber types for opioid prescribers. Additionally, reviewing total opioid prescribers grouped by prescriber type helped find an anomaly that could be profiting from the opioid crisis.

**Communicate:**The team prepared a report and presented processes, findings, model techniques, and accuracies in class. Key takeaways involved discussing the anomaly, high-risk States, and high-risk prescriber types. For example, Family Practice has one of the highest conditional probabilities for opioid prescriber = True, and, unfortunately, most of us see a general doctor at a Family Practice. Moreover, the anomaly identified requires further investigation.

**Lessons Learned:** The project encourages communication in a data science team, tuning parameters to improve models, and re-enforced the importance of a hold-out test vs. cross-validation. At one point, the code failed during predictions due to unbalanced training and testing sets. Our results highlight the importance of normalization; largely populated States resulted in high conditional probability values for the opioid prescriber = True and opioid prescriber = False labels. Lastly, these models can take a long time to run. The opioid analysis project emphasized the significance of considering cloud servers for building Machine Learning models, as explored in Cloud Management class.

The Text Mining course expanded the Data Science toolbelt by teaching to apply visualization and predictive modeling techniques to unstructured data to help generate actionable insights. The team selected a Wikipedia Comments corpus[[4]](#footnote-4) to identify toxic comments, which turned out to be a challenging and sensitive topic. Developing word clouds of features generated after corpus vectorization helped remove unhelpful features before training Support Vector Machines, or SVM, models to classify comments as toxic or not toxic. Unhelpful features became apparent as features that did not help differentiate between toxic and non-toxic comments. After training the SVM models, bar plots helped reveal the most indicative words used to classify comments as toxic or not toxic.

The Wikipedia Toxic Comments analysis revealed the responsibility held by the Data Scientist to be aware of bias naturally built into datasets. Before building the SVM models, removing features relating to cultures, ethnicities, nationalities, governments, sexual orientations, and any racial slurs was crucial. The goal was to avoid the model labeling a comment as toxic based on the existence and frequency of a word referencing a culture or ethnicity. One of the SVM models trained produced a list of indicative words for toxic comments consisting of the word "Khoi." The Wikipedia Comments corpus uses "Khoi" as a sign-off for comments. Curiosity led to a google search, and the first definition retrieved showed that Khoi is a word used to describe a group of Indigenous people. It would be erroneous for an SVM model to use a word like Khoi to label toxic comments. Bringing in a subject matter expert to help detect obstructive words during feature engineering may help avoid model bias.

The Applied Data Science Master of Science program at Syracuse University facilitated my growth as a Data Scientist. A mix of the Program Learning Goals is implemented in all courses, ensuring a well-rounded experience. The relevant coursework boosts confidence when interviewing for roles. The program elevated my Data Science skills, allowed easy connections to fantastic people, and improved life quality.

1. 1. Property by Ashley Database – IST 659 Link:

   <https://sumailsyr-my.sharepoint.com/:f:/g/personal/jmlogron_syr_edu/EtcG_-jA20FKqnl1P_UYe1YBIBrkr93nP5isSRXosn00jQ?e=n4pHZG> [↑](#footnote-ref-1)
2. 2. Wildfire Danger Zones in California – IST 719 Link:

   <https://sumailsyr-my.sharepoint.com/:f:/g/personal/jmlogron_syr_edu/EkAOVnBYR8pMqb_7WEfLjJ0BnM1iP4V3Wc36FYas7XEAwg?e=VAec82> [↑](#footnote-ref-2)
3. 3. Opioid Prescribers Analysis – IST 707 Link:

   <https://sumailsyr-my.sharepoint.com/:f:/g/personal/jmlogron_syr_edu/Eqmnb78CIRVJor6uIRBo-UkBRlZaftac1iu8WaIEkwCUUQ> [↑](#footnote-ref-3)
4. 4. Wikipedia Toxic Comments Analysis – IST 736 Link:

   <https://sumailsyr-my.sharepoint.com/:f:/g/personal/jmlogron_syr_edu/Ev7zGncS1EhNsGW57DZworUBb1chAOfHHLSkENHFdLZARg?e=hXWwwv> [↑](#footnote-ref-4)