**Introduction**

Syracuse University’s Master of Science in Applied Data Science program is interdisciplinary curriculum focuses on delivering organizational insight and driving business strategy by using data capture, management, mining and analysis skills. The main goal of the program is provided a mixture of technical and theoretical skills with for today’s professionals. The program focuses on data science but includes an understanding of how to handle data, how to capture data as well as how to manage and analyze data in order to make informed decisions.

The portfolio reflects on 4 projects that were completed during the programs that focuses on the following 7 Key Learning Goals of the Program –

1. Describe a broad overview of the **major practice areas** of data science.
2. **Collect and organize** data.
3. **Identify patterns** in data via visualization, statistical analysis, and data mining.
4. **Develop alternative strategies** based on the data.
5. Develop a plan of action to **implement the business decisions** derived from the analyses.
6. **Demonstrate communication skills** regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization.
7. Synthesize the **ethical dimensions** of data science practice (e.g., privacy).

**Project 1 : IST 736 - Text Mining**

**Project Goal**

|  |  |
| --- | --- |
| Although William Shakespeare is the most-recognized  playwright in the world, very little of his life is known. This has  led to many conspiracy theories that suggest that Shakespeare  may have been more than one person. |  |

The goal of the project was to analyze Shakespeare’s plays using **Text Mining** to see if there was evidence that it was written by more than one person and to see if there was any linguistic difference between the Male and Female characters in his plays.

**Data collection and Preprocessing**

A few plays from each of the Categories “Comedy”, History” and “Tragedy” were chosen to form a collection of **15 plays** that were used for the analysis. Word Cloud indicates that the word “**love**” featured as prominent topic in many of his plays. Other words like “**sir**” and “**thou**” were used commonly in his plays.

|  |  |
| --- | --- |
|  |  |

**For the Gender Analysis,** a corpus with 9000 speaking lines was created using Python Code with about 4500 for Female Character and 4500 for Male Characters so that the corpus was equally balanced between the genders. Screen shot below show the dataframe that was created for Gender Analysis -

**Standard pre-preprocessing** steps like converting the words to lower case and removing punctuations were applied. When reviewing the dataset, it was clear that NLTK’s English Stop words will not be enough since the dialogues are written in what is called “Early Modern English” and would require some additional stop words to remove words like “thou”, “yonder”, “art” etc.



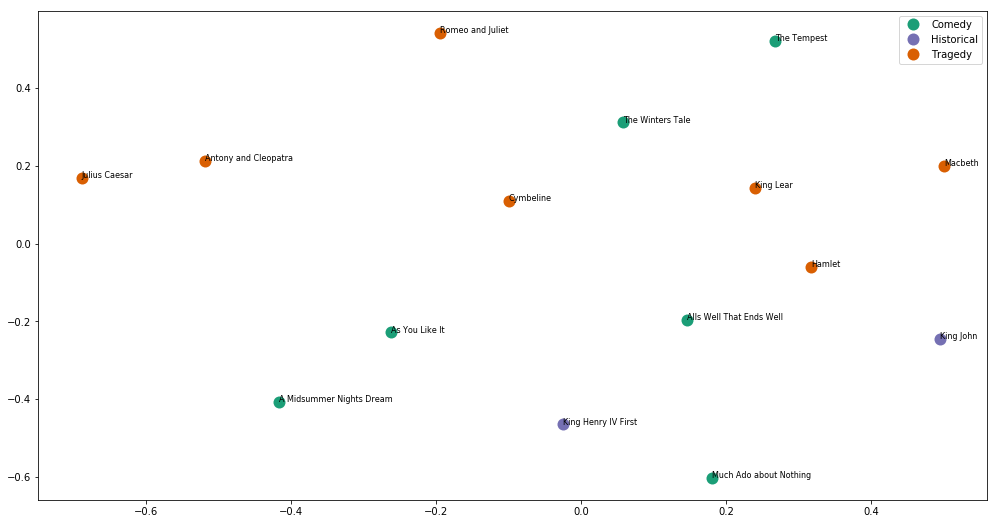
A **custom list of stop words** for “Early Modern English” was created and appended to NLTK’s English Stop words which resolved the issue



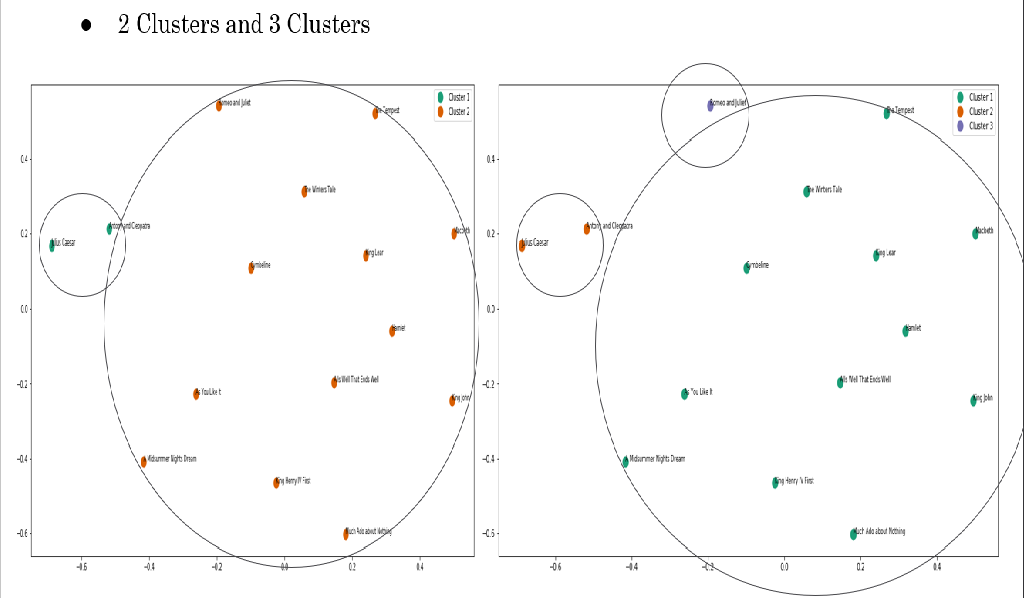
In order to analyze the **conspiracy theory** that Shakespeare’s works may have be written by other people, a selection of poems attributed to William Shakespeare was compared against poems attributed to both Christopher Marlowe and Edward de Vere. In total there were **52 poems** were analyzed using clustering techniques, with 19 from Shakespeare, 20 from Edward de Vere, and 13 from Marlowe. The reason for using only poems is to compare similarly styled writings, as plays and poems could lead to potential clustering issues between the authors. Furthermore, they were the only style of writings found for the possible alternate authors. Finally, due to the highly artistic style of poetry very little preprocessing was done. Only punctuation marks were removed as they were mostly artifacts from the formatting of the poetry.

**Data analysis and Visualization**

Plotting each play by **Euclidean distances** gives the first glance that it is plausible that more than one person authored these plays. Typically, works from a given author should be closely grouped as they tend to use the same writing style.



Further analysis using plots with various **clustering options (K-Means)** indicates that it is most likely that **two or three authors** wrote these plays. Any more than that is improbable with the current analysis.



In order to evaluate the **feasibility of performing a Gender Analysis** on Shakespeare’s Characters the Word Frequency Plot for the 20 most used words for each Gender was created and analyzed to see if any of the common words was used disproportionately by one gender. As the screen shot below indicates, there is some **gender-imbalance in the** words used in the dialogues as indicated by the plot.

|  |  |
| --- | --- |
|  |  |

Using **Matplotlib**’s **pyplot** function a bar chart of the top 20 words was plotted. Looking more closer to the top 20 words used by both genders, it is clear that there are some words that were used commonly by both Genders like “**love**”, but the frequency however was very different. The word “**love**” was used in over **150 speaking lines** for **Female** characters in the Dataset but only about **100 speaking** lines for **Male** characters. In addition, there are words like “**death**” the was commonly used in Male speaking lines but not so much in Female speaking lines.

|  |  |
| --- | --- |
|  |  |

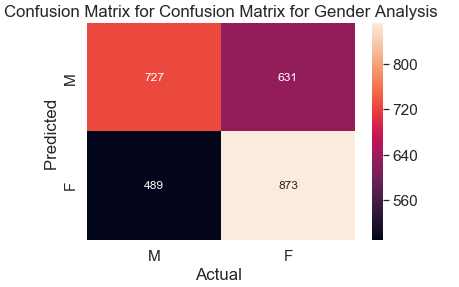
The analysis clearly suggests some distinction in the dialogues between the Male and Female characters. The next step was to see if commonly used Text Mining Algorithms can be trained to distinguish these differences.

**Text Mining Algorithms**

The Dataset was split using the Hold-Out method to use 70% for Training and 30% for Testing. Three vectorization options were analyzed, the first was using an N-gram Count Vectorizer, the second using an N-gram Boolean Vectorizer and a third using a N-gram TfidfVectorizer with min\_df set to use .0001%. Although both SVM and MNB algorithms rendered similar results for the different vectorization options, MNB consistently scored higher compared to SVM with MNBusing **N-gram Count Vectorizer** scoring the highest Accuracy of **59%.**

|  |  |  |
| --- | --- | --- |
|  |  |  |

Looking at the Confusion Matrix for the **MNB using N-gram Count Vectorizer**, 631 speaking line that belonged to the Male Character were wrongly predicted as Female character and about 489 Female speaking lines were also wrongly predicted as Male.



To build a **case for alternate authors,** multiple models using different methods, parameters, and seeds would be used to compare their consistency in predicting each other. The two main model types used were **support vector machine (SVM) and multinomial Naive Bayes models (MNB).** The types of parameters used for each model were unigram only Boolean, unigram only term frequency (TF), and unigram + bigram TF. Finally, a total of 7 different seeds were used.

|  |  |  |  |
| --- | --- | --- | --- |
| Confusion Matrix Average | Christopher Marlowe | Edward de Vere | William Shakespeare |
| Christopher Marlowe | 1.19047619 | 0.380952381 | 2.142857143 |
| Edward de Vere | 0.785714286 | 4.071428571 | 1.571428571 |
| William Shakespeare | 2.547619048 | 0.642857143 | 2.642857143 |

When comparing the confusion matrices across all the models and seeds, one thing that stood out was the difficulty in the models differentiating between Shakespeare and Marlowe. Looking at the confusion matrix average across all models, the average misclassification between **Marlowe** and **Shakespeare** stands out as the **most misclassified poems**. The misclassifications averaged over 2 poems either way between the two when the test sets consist of 16 total poems. The only other misclassification of note is classifying Edward de Vere’s poems as Shakespeare’s poems at roughly 1.5 poems out of 16. Interestingly though, Shakespeare’s poems were not frequently classified as de Vere’s poems.

To further build evidence for potential alternate authors of William Shakespeare, the conditional probabilities of the MNB models would be compared.



The probabilities would be distributed into 6 potential categories. For each model, they would be classified based on the authors that had over 10% likelihood predicted. Based on the counts, the models were good at classifying de Vere’s and Shakespeare’s poems but not Marlowe’s. Marlowe’s poems instead appeared to be confused with Shakespeare’s poems or Shakespeare’s and de Vere’s poems, but not solely with de Vere’s poems**. This is just further evidence that Marlowe could be a potential alternate author for Shakespeare**.

**Reflection and Leaning Goals**

Text mining algorithms are as good as the data that is fed to the algorithm - as the saying goes garbage in is garbage out. Data gathering and preprocessing are the most essential and time-consuming step in text mining. It is said that data scientists spend 80% of their time cleaning and preprocessing data and only 20% of their time analyzing. Investing the time to collect the right data, understand the data and applying the correct pre-processing steps is key to successful text mining!

The choice of Algorithms for text mining depends on the Data being analyzed as well as the type of Classification and is clearly an important decision since different Algorithms give different accuracy results. But equally important is the Quality of Data and the Correct Vectorization options. The take away is that the choice of the Algorithm and Vectorization Options depend entirely on both the Data and the Classification and are key to text mining!!

Several learning goals such as learning to **collect** the necessary data, applying standard preprocessing techniques to clean and **organize** the data, **visualize** and analyze the data via Word Clouds and Word Frequency Distribution Plots, applying multiple **text mining algorithms** such as SVM and Naïve Bayes and interpreting the results to gain **meaning insights** from the results.

**Project 2 : IST 565 - Data Mining**

**Project Goal**

**Black Friday** is an informal name for the Friday following [Thanksgiving Day in the United States](https://en.wikipedia.org/wiki/Thanksgiving_(United_States)), which is celebrated on the fourth Thursday of November and is the beginning of America's [Christmas shopping season](https://en.wikipedia.org/wiki/Economics_of_Christmas) since 1952. Many stores offer highly promoted sales on Black Friday and open very early, such as at midnight, or may even start their sales at some time on Thanksgiving.

The goal of the project was to use data mining to analyze a sample dataset from a retail store to who the best customers are, what pushes them to shop, how frequently they buy, how much they spend per order and train models to predict the Gender of the Customers for future Black Friday Sales.

**Data collection and Preprocessing**

The “ Black Friday Sales Data “ from Kaggle ([https://www.kaggle.com/mehdidag/black-friday#BlackFriday.csv](https://www.kaggle.com/mehdidag/black-friday)) was chosen for the project because it is a sample retail transactions containing a variety of data ( discrete and continuous) and offers an opportunity to apply various data mining algorithms to analyze the data. The data set is large with 12 variable and about 530K observations containing the purchase pattern for buyers at a retail store - what products were bought and how much was spent. The data was analyzed for missing data and duplicates and found to be a relatively clean dataset.

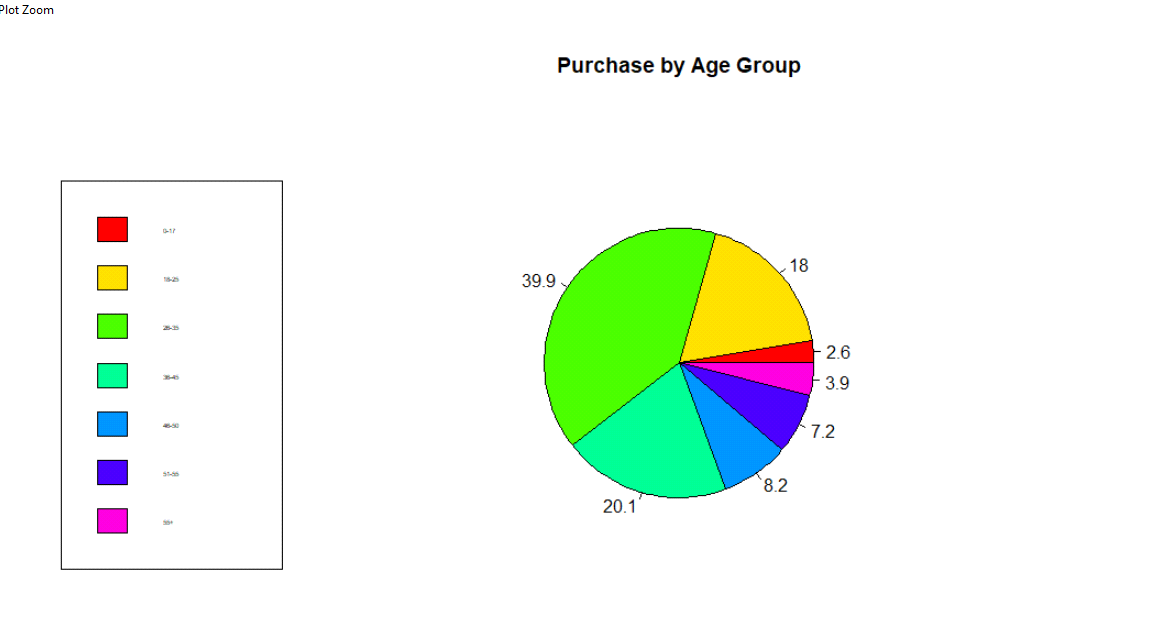
Various Pre-processing options were applied depending on the algorithm implemented.

**Data analysis and Visualization**

Visualizing data can give interesting nuggets –

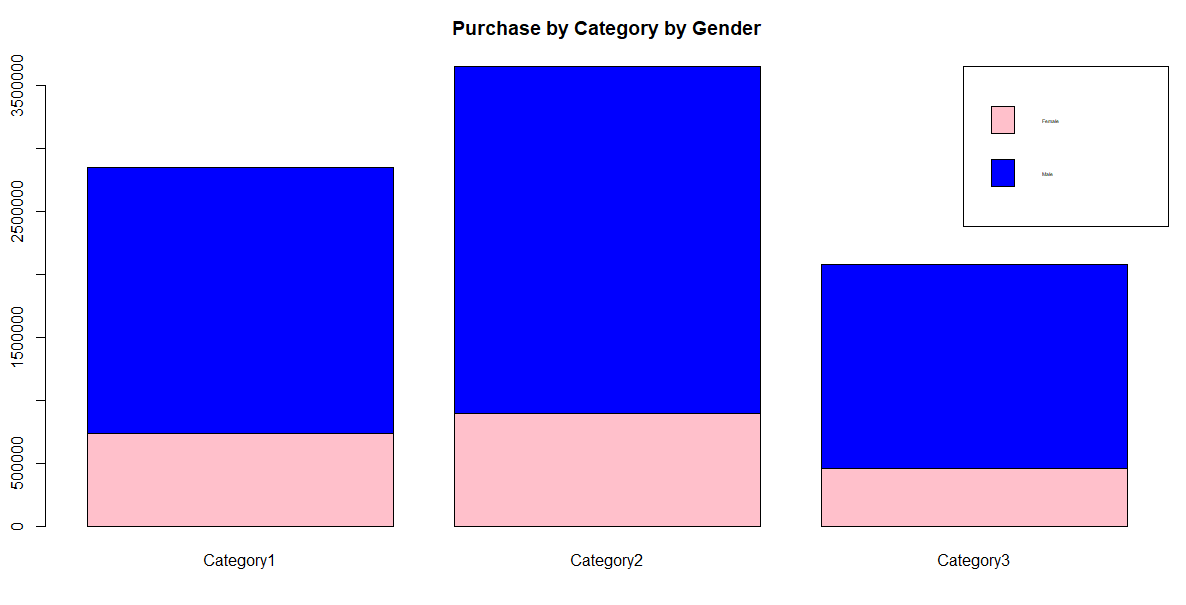
**Purchase habits based on the Age Bracket:** The majority of the buyer were in the Age Bracket 36-45 followed by people within the Age Bracket 26-35. This makes sense because these are age groups that employed and have spending power.

**Age**



**Purchase habits based on Gender:** The bar chart analysis for Gender also shows spending habits on Men versus Women. Male buyers were clearly the biggest spender which is a little surprising because people generally think that Women are big on shopping! This may be due to Online shopping where Men tend to dominate compared to Women. Also, the Category 2 which is “Electronics” had the greatest number of buyers which is not a surprise since most buyers were Men.

**Gender**

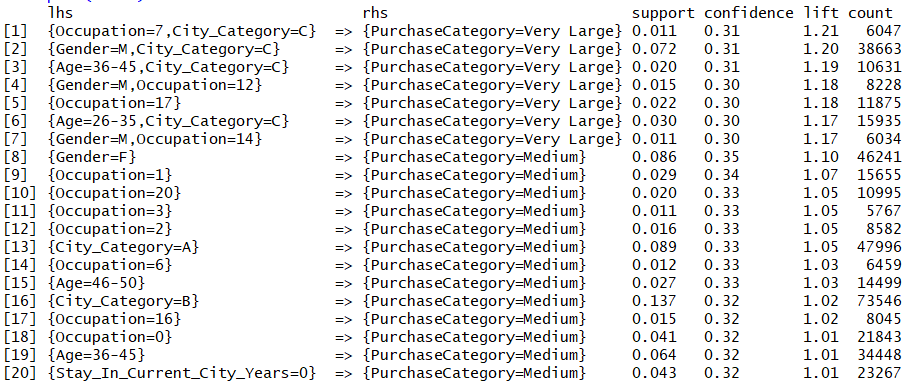


**Data Mining Algorithms**

**Association Rules Mining**

Association Mining searches for frequent items in the data-set and is an excellent algorithm to study the frequent item sets along with the association rule. The Apriori Algorithm was used to get meaningful insights into the association rules for purchase categories.

Association Rule Mining was used as Supervised Learning Method by forcing the target calcification attribute “Purchase Category” on the RHS of association rules to find out which attributes of the dataset contribute to a given Purchase Category. After removing redundant rules, the algorithm returned 35 rules



A few insights from these rules-

* Rules 2, 4 and 7 indicate that Male shopper have a high probability of making Large Purchases ($25000 - $12000) whereas Rule 8 indicates that Female buyers tend to make medium purchases ($5000 - $ 8000).
* Rules 1,2,3 and 6 show that buyers that live in City Category “C” tend to make Large Purchases indicating that it may be an affluent City where as Rule 13 indicated that buyers from City Category “A” are more likely to make medium purchases.
* Rules 3 and 6 indicate that buyer in the Age Group of 26-35 and 36-45 are more likely to make Large purchases where as buyers in the Age Group 45-50 (Rule 15) are more likely to make Medium Purchases.

**Classification**

Four classification Algorithms - Random Forest, Naïve Bayes, knn and Support Vector Machine were trained using various parameters **predict the Gender of the buyer.**

The summary of the data mining algorithms clearly shows the data and its feature like high-dimensionality, sparsity, multi-view, multi-label define which algorithm will be suitable. Classifier that work well with a given dataset may perform poorly with another dataset.

|  |  |  |
| --- | --- | --- |
| **Model** | **Parameter** | **Accuracy** |
| **Random Forest** | Mtry, Ntree | 85.39% |
| **Naïve Bayes** | Usekernel, Adjust, fL | 75.25% |
| **knn** | k | 84.55% |
| **Linear SVM** | C | 75.70% |
| **Radial SVM** | C, Sigma | 75.13% |

**Reflection and Leaning Goals**

Nothing is more important to a retailer than really knowing their customers. With data mining algorithms they can learn exactly who their best customers are, what pushes them to shop, how frequently they buy, how much they spend per order, and more.

Since the dataset had transactional information regarding purchases at a retail store, the Association Rules Mining proved to be very value data mining tool since it could identify some important rules such as the population of buyer that make Large or Medium or Very Large purchases. For a retail store, this analysis is very valuable in determining who are the big item buyers, what age bracket do they belong and where do they live.

Retailer are interested in analyzing their buyer to see who they should target for future marketing and advertising. The Data Mining Algorithm were able to predict the Gender of the Customer with a relatively high accuracy of 85%. This will help with better marketing and advertising strategies to improve future black Friday sales.

This project provided an opportunity to demonstrated several leaning goals such as **collect** the necessary data, applying standard preprocessing techniques to clean and **organize** the data, **visualize** the data using pie charts and bar charts , applying several **data mining algorithms** like Random Forest, knn, Naïve Bayes and SVM to analyze their predictive performance and accuracy against the Black Friday Sales Data.

**Project 3: IST 659 - Data Administration**

**Project Goal**

The goal of the project was to build a database for a Karate Studio database to access and maintain student and employee information and track class schedules for all the locations which include -

1. Track general information about their studio locations like the address, studio manager etc.
2. Track and maintain information about all their employees like address, salary and employee schedule.
3. Class Schedule for their studio locations.
4. Track and maintain information about the students like name, address and class they are currently registered for.
5. Event Schedule for their studio locations.

**Database Architecture**

The first step to building a database is to collect the necessary **business rules** which are as follows-

1. Each studio location has one or more of classes and a class can be held in one of more studio locations.
2. A class can have multiple sessions and each session is for one class only.
3. Each session can have one or more students and a student can be register for only one session at any given time.
4. A student may have a Parent (An Adult Student will not have Parent Information) and a Parent must have one or many Students.
5. Studio employees can be full-time, part-time and apprentices. An employee can work in one or more studio location and a studio location can have one of more employees.
6. Salaries for the employee also various depending on the employee type.
7. A studio can have multiple events scheduled and an event can be scheduled for one or more studio locations.

The next step was to design the Conceptual ERD and Logical ERD models

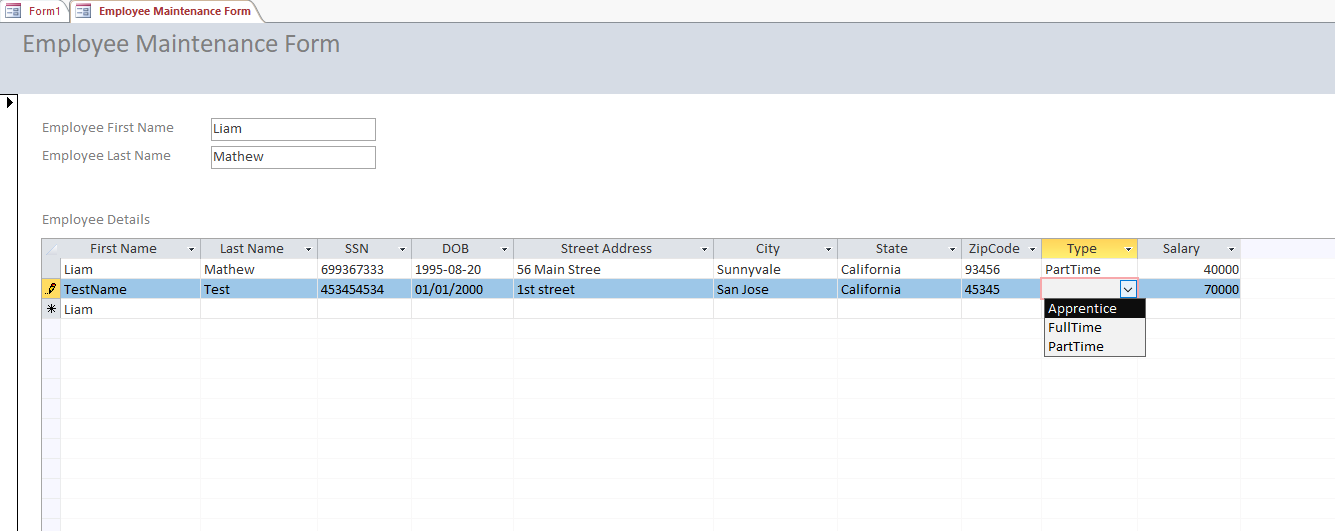
The following is the Normalized ERD model that was designed for the database

****

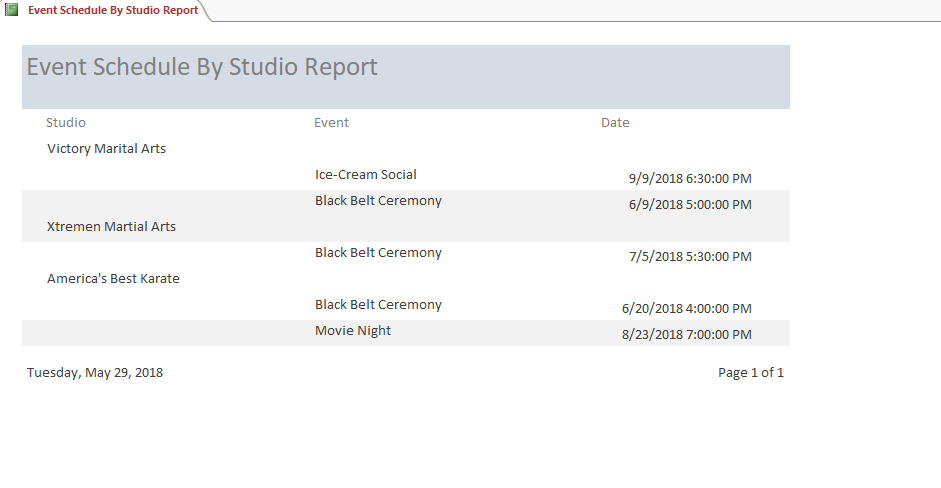
**Implementation**

**MS SQL** was used to implement the Database Tables and Views. The Business Rules were implemented using Stored Procedure and Functions. For example, a Stored Procedure “sp\_AddModify\_Employee\_Info” was developed to imaintain information about employees like address, salary and employee schedule and allows for adding and updating employee information.

**MS Access** was used to implement the Data Entry/Maintenance screens and Reports. Screen shot below is the MS Form to view existing employee as well as add new employees

****

Sample Report

****

**Reflection and Leaning Goals**

Data is the most important asset to any Data Scientist. Data modeling is a crucial skill for every data scientist. The ability to think clearly and systematically about the key data points to be stored and retrieved, and how they should be grouped and related, is what the data modeling component of data science is all about.

Data scientists and analysts deal with complex data. In order to make use of this data, significant effort is spent in data engineering. Data engineering transforms and normalizes complex data into relational databases or into an output format that can then be loaded into data science tools to derive insights.

This project demonstrated the importance of investing enough time to design the conceptual and normalized data model which is the foundation to any database. The database design needs to be flexible so that when business rules change, the design can be updated with minimal changes to the overall database.

Several key goals were demonstrated like **collecting data** and **organizing the data** into tables, **understanding the business rules** to implement the business logic via stored procedures or functions and **analyzing and visualizing** the data using Reports.