Portfolio Milestone Report

Jonathan Durbin | SUID 510804910 | jdurbin@syr.edu

Towards completion of the
Master's Degree in Applied Data Science
Syracuse University School of Information Studies

Table of Contents

Introduction

In an era defined by the explosion of data and the increasing demand for informed decision-making, the field of data science has emerged as a force driving innovation across many domains. This paper describes a journey undertaken within the framework of an Applied Data Science Master's program. It showcases the attainment and understanding of several key objectives that lie at the heart of the discipline. The goal of this endeavor is to master the art of collecting, storing, accessing, and analyzing data while harnessing the power of relevant technologies. Through rigorous training and practical exposure, this paper demonstrates the application of technology to build robust data solutions that serve as the foundation for insightful analysis. Such applications range in topic from movies to police departments; from databases to natural language processing. Central to these projects is the proficient use of programming languages, particularly R and Python, which serve as the backbone of data manipulation, analysis, and model development. The journey described in this paper extends beyond technical prowess - the ability to effectively communicate findings to a diverse audience is equally important. This paper highlights the art of writing compelling narratives through visualization and analytics.

The Applied Data Science program has a number of learning objectives that this paper will demonstrate. By the end of this paper, it should be clear that the student is able to:

1. Collect, store, and access data by identifying and leveraging applicable technologies.
2. Create actionable insight across a range of contexts, using data and the data science life cycle.
3. Apply visualization and predictive models to help generate actionable insight.
4. Use programming languages such as R and Python to support the generation of actionable insight.
5. Communicate insights gained via visualization and analytics to abroad range of audiences.
6. Apply ethics in the development, use and evaluation of data and predictive models.

IST 652 Scripting for Data Analysis

Project Description

This was a straightforward project – the goal was to perform some data analysis on six datasets containing information about movies and TV shows as they are rated on different entertainment platforms (Amazon prime, Disney+, HBO Max, Hulu TV, Netflix, and Paramount). For each platform, there are two datasets. The first contains a list of actor and director credits for all titles. The second is a list of titles and their ratings. The data was first cleaned using Python regular expressions then saved in a separate location. Analysis in this project consisted of answering a number of data questions, such as "What is the average runtime per platform?" (*Figure 1*) and "What is the average TMDB and IMDB score for each platform?" (*Figure 2*). From the graphics created with these questions and more, a presentation was created that explored the dataset and the insights that were previously gathered.

The project was completed in a group of two. Work was evenly and fairly split between developing a Python notebook and working on the final write-up / presentation.

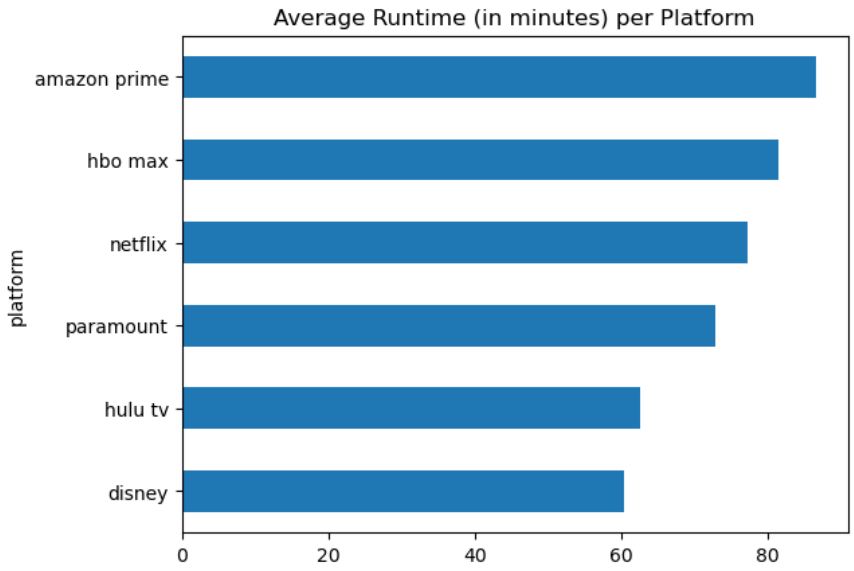

Figure 1: Avg. runtime per entertainment platform

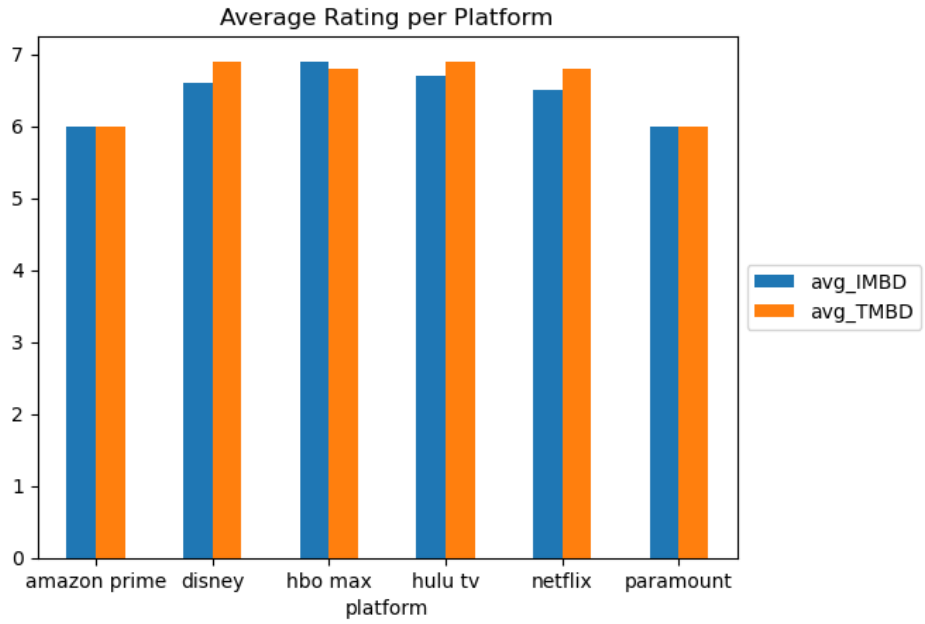

Figure 2: Avg. rating per entertainment platform

Reflection and Learning Goals

By engaging in an exploration of six datasets, each offering insights into movie and TV show ratings across prominent entertainment platforms, this project demonstrated competency in various data science skills. The process of data collection and cleaning, facilitated by the extraction of information from online sources and the development of a pipeline to clean the data, underscored the adeptness in identifying and leveraging relevant technologies. Following the stages of data analysis – coming up with a data question and finding a way to answer it – is a testament to the application of the data science lifecycle in the real world.

IST 659 Database Administration Concepts and Database Management

Project Description

This project was less about analyzing data and more about developing a way to efficiently organize and query the data. The goal of this project was to build a database for a fictional police department, populate it with fake data, then build an application that could both ask questions of the database and push updates to it. The first part of this project was greatly aided by a phone interview with a local police department clerk. After this interview, the number of tables in the database and how to connect them together was a little bit clearer. The interview revealed that the Auburn Police Department (APD) has four primary tables – Names, Warrants/Wants, Reports, and Complaints/Incidents.

As a part of the submission process, a list of business rules was created. They aided in the design of the database by placing constraints on how entities within the police department relate to each other. It is likely that these constraints differ slightly from the real-world constraints used in the APD database. An example of such a rule is "A **person** is defined as someone who has been in some form of contact with the APD (or some fictional police department). Officers, suspects, victims, etc. are on this list." Stakeholders also were identified as a part of the submission process. In the case of this project, stakeholders would be those who work for the APD - Officers, Detectives, clerks, and other city officials.

A conceptual model of the database was constructed (Figure *3*), then normalized to a logical model (Figure *4*). From there, the physical design of the database took place – writing SQL scripts to build tables that reflected the logical model. Once the tables were built, they were populated with fake data. Finally, several functions, views, and procedures were created. The functions and views answered possible data questions an officer or clerk might have, while the procedures allow the addition or removal of information from the database.

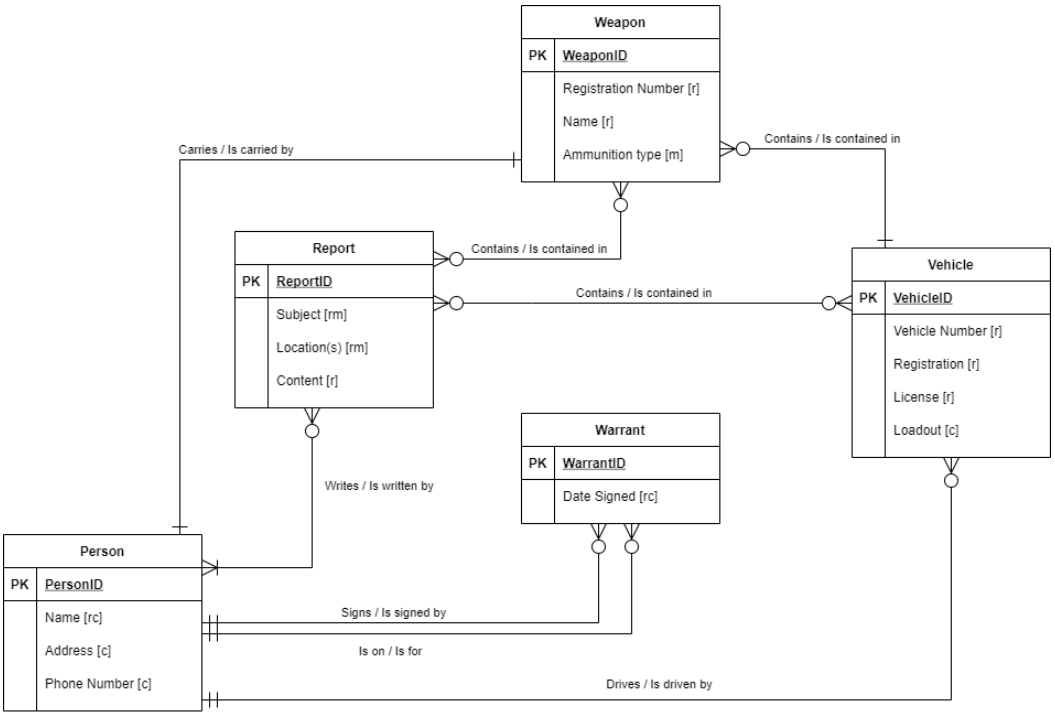

Figure 3: A conceptual diagram for a fictional police department's private database.

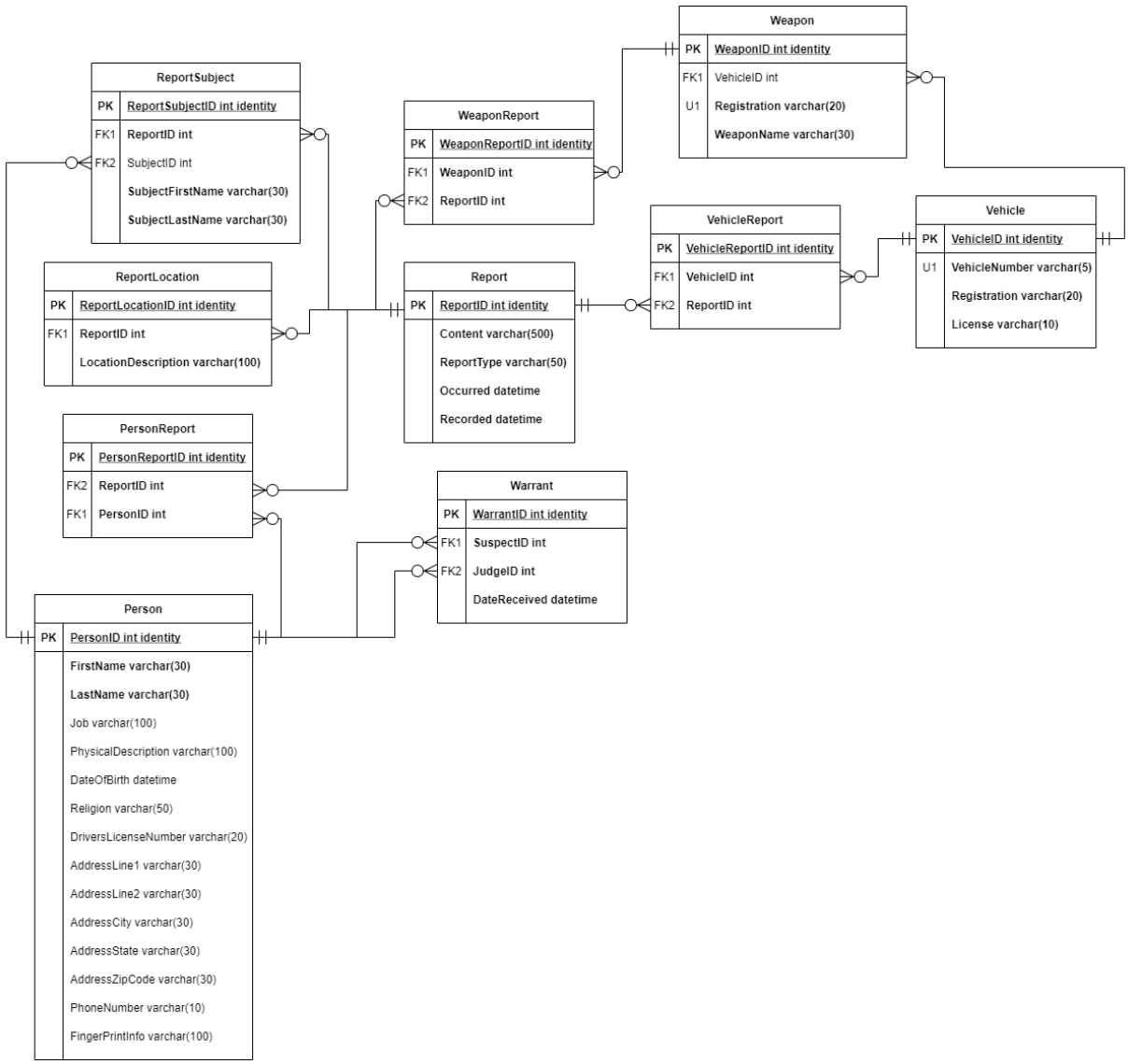

Figure 4: A normalized logical model for a fictional police department's private database.

Once database development was finalized, a simple web application was created using the Shiny package in R. It features a simple, proof-of-concept interface that allows toggling between the different views (*Figure 5*), an update form (*Figure 6*), and the more specific data questions (*Figure 7*) (for the purpose of allowing the professor of the class to quickly see that those specific questions were answerable).

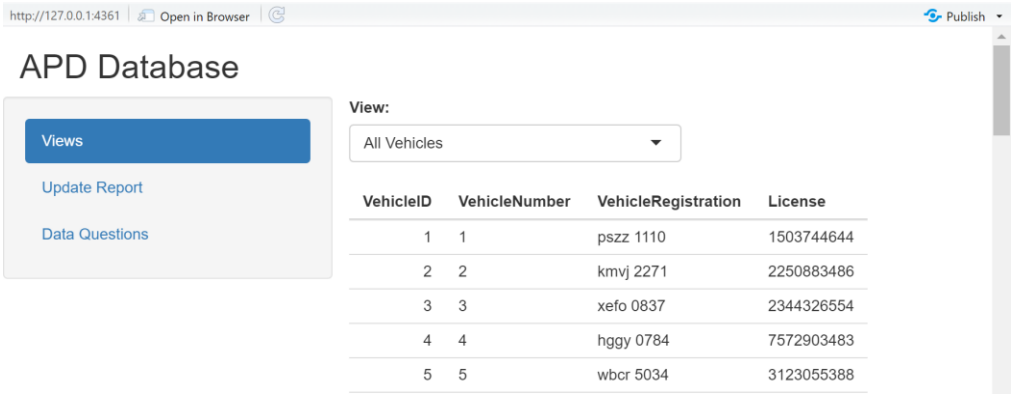

Figure 5: An example of one of the views for the R Shiny web application.

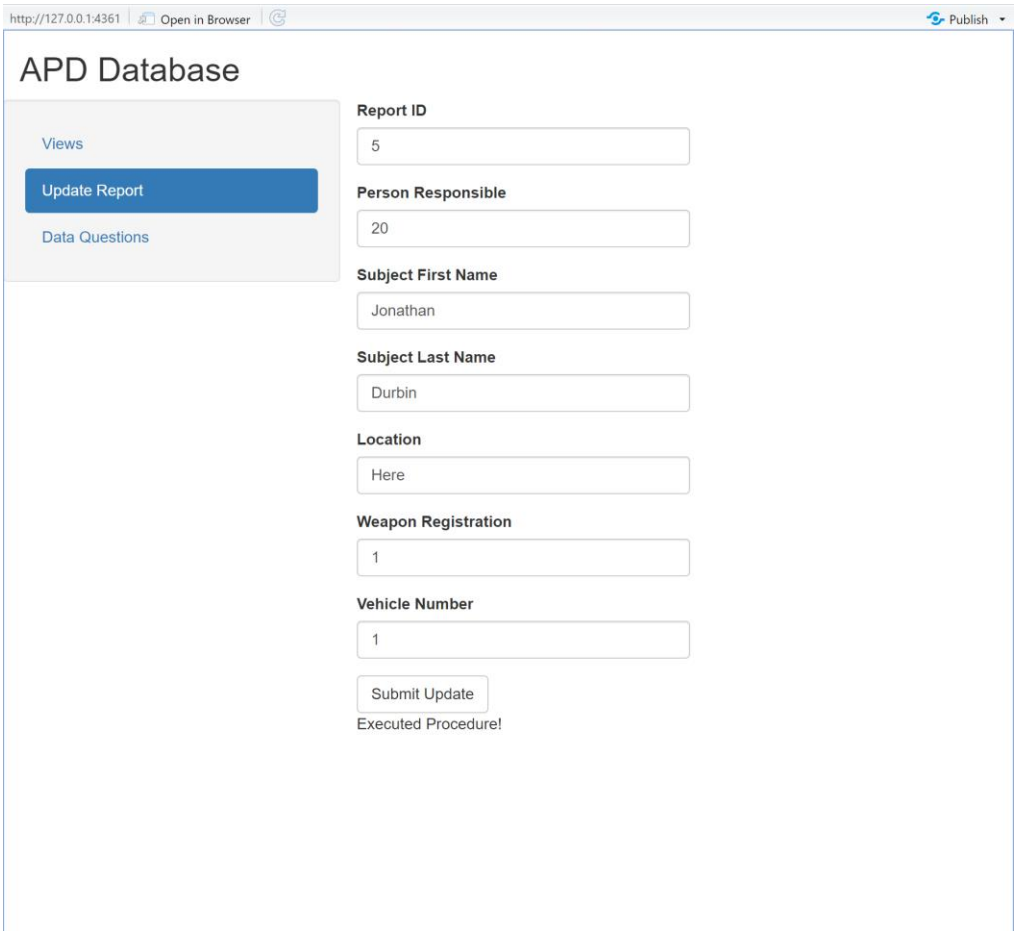

Figure 6: An example of the update form on the R Shiny web application.

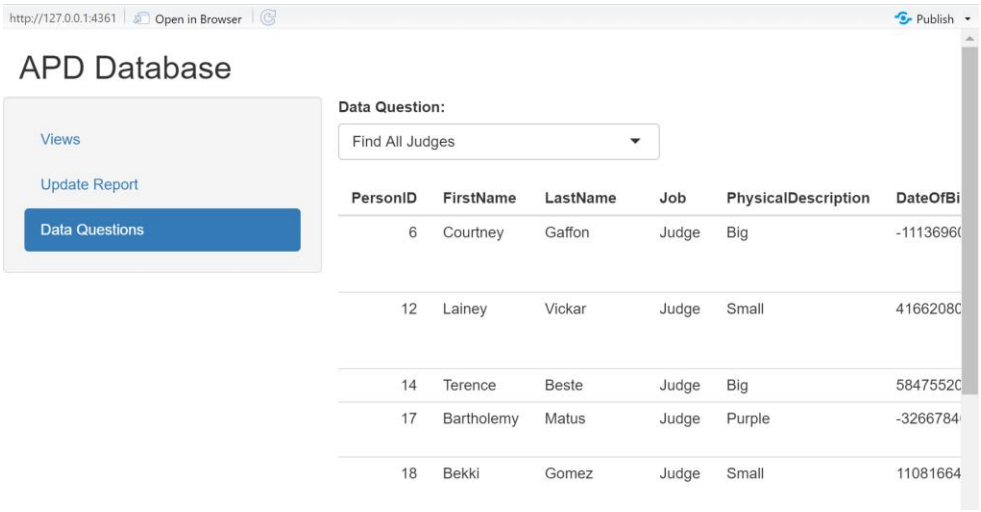

Figure 7: An example of one of the data questions on the R Shiny web application.

Reflection and Learning Goals

This project served as a demonstration of the overarching learning objectives within the Applied Data Science Master's program. By conceptualizing, normalizing, and subsequently implementing a database structure using SQL scripts, the student was able to gain a deeper understanding of how to use technologies that are relevant to the industry today. The act of transforming a conceptual model into a logical and then physical design showcased a thorough understanding of database architecture, which is a demonstration of the underlying goals of collecting, storing, and accessing data effectively. Furthermore, the project's emphasis on developing a comprehensive database solution aligns with the objective of generating actionable insights across a range of contexts. Being able to efficiently organize and query data reflects the real-world needs of police departments across the world. The incorporation of functions, views, and procedures demonstrates the ability to translate database architecture into practical tools that officers and clerks could utilize to retrieve information and enact updates, which fulfills the goal of actionable insight generation. Through this intricate orchestration of technological proficiency and solution development, this project mirrored the program's learning goals.

IST 707 Data Analytics

Project Description

The goal of this project was to solve a real-world data mining problem. A problem was defined in terms of its practical business application. In this case, an attempt was made to perform association rule mining and text mining on some recipe data. The inspiration to seek out recipe data came from the fact that there is no major recipe database in existence and that it is an interesting challenge to think about how best to talk about recipes at a low-level layer of abstraction (i.e. in a programming or query language). The recipe data was gathered from eightportions.com, where it is split into three groups – recipes from Foodnetwork, Epicurious, and Allrecipes. About 70,000 of the recipes had images attached to them, but these images were ignored as they were not relevant to the analysis. After some basic data cleaning, the data was pared down to only ingredient names (e.g. converting "1 ½ cups sugar" to "sugar") using a Python package.

The apriori algorithm in the Python arules package was used to perform the association rule mining analysis. Text mining consisted of converting the list of recipes into a corpus (a collection of words), storing that as a list of bigrams, which are then rated according to a measure called Pointwise Mutual Information (PMI). The association rule mining analysis found that in the top 20 association rules, all-purpose flour is the most used ingredient (*Figure 8*). Meanwhile, the text mining analysis found that alcoholic drinks are a common occurrence in the dataset. As seen below, most of the bigrams are names of vineyards (*Figure 9*).

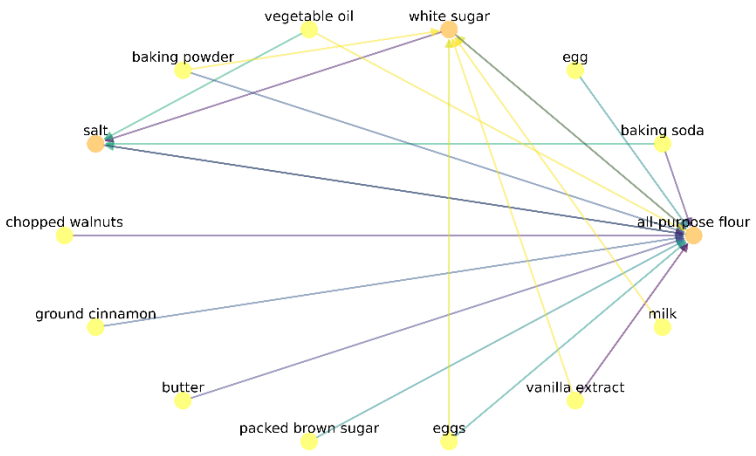

Figure 8: A diagram of an association rules network.

```
('bedford', 'thompson'): 21.1077
('coyote', 'bait'): 21.1077
('hahn', 'estates'): 21.1077
('nitrous', 'oxide'): 21.1077
('sabor', 'italia'): 21.1077
('kung', 'pao'): 20.8446
('vice', 'versa'): 20.8446
('anna', 'williams'): 20.6222
('frankland', 'estate'): 20.6222
('gale', 'gand'): 20.6222
('grana', 'padano'): 20.6222
('paula', 'deen'): 20.5816
('nam', 'pla'): 20.4296
('kent', 'goldings'): 20.4296
('bungee', 'cords'): 20.2597
('piment', "d'espelette"): 20.2597
('pregnant', 'women'): 20.1077
('ina', 'garten'): 19.9702
('buon', 'appetito'): 19.9557
('bagna', 'cauda'): 19.8446
```

Figure 9: A list of bigrams from the Foodnetwork corpus, ranked by PMI.

Reflection and Learning Goals

The exploration of association rule analysis and text mining within the context of a recipe dataset demonstrates a union of data science skills, aligning with the learning goals within the Applied Data Science Master's program. The goal of this project – tackling a genuine data mining problem – reflects the practical and business-centric orientation of the program's objectives. The endeavor to unravel relationships between ingredients in recipes using association rule mining showcases the application of data science techniques to real-world scenarios. Additionally, the project's venture into text mining embodies the art and skill of transforming unstructured data into valuable insights.