Syracuse University

School of Information Studies

M.S. Applied Data Science

# *PORTFOLIO MILESTONE*

Mark A. Roberts

SUID: 598273961

Github: https://github.com/dodekete/MSADS_Portfolio

## *MY INTRODUCTION TO DATA SCIENCE*

A career in marketing/business development design and production professional in the commercial real estate industry served as my introduction to data science. I worked for a premier office and industrial real estate firms in the Greater Boston market. My firm specialized in biotech/life science but also high-end office and industrial leasing assignments for firms like TJX, State Street, Converse, Harvard Medical School/Boston Children's Hospital/Dana-Farber Cancer Institute, Corning, Boston Consulting Group, Shire Pharmaceuticals, TA Associates, Invesco, TIAA-Cref, Brookfield Properties and Normandy to name a few. Corporate real estate occupiers and owners had relied somewhat on power of relationships; working with a broker who was your college roommate whose insight was trusted implicitly. It was a 'Boys' Club'. After the Dot.com bust in 2001 and the banking crisis of 2008, that pattern rapidly became archaic. Suddenly, users needed "justification" for making significant moves, like moving a headquarters or leasing multiple stories of a downtown office. Enter data science.

Now, stakeholders were more diverse, not likely to be impressed by box seats at the big game, needed evidence to support significant transactions. Space utilization studies, employee commute analysis, labor supply dynamics, NER effects and lease up assumptions. Many of these items simply served an anecdotal and often purely visual representation that the move to a brand-new facility was not only justified, but necessary. There was a limit though: nobody wanted a regression equation, just a nicely color coded map to illustrate a(n already arrived at) decision point. I was keenly aware that we weren't "proving" anything but rather presenting a view of data that affirmed a conclusion we had already reached. Data was there to be there, not to be intrinsic to business practices. Still, it was better than nothing.

## *PRECONCEIVED NOTIONS*

From watching too many procedural and fantasy/sci-fi shows, I was abused of the notion that data science was a magic trick. Pour in the right data, parameters and voila, an answer emerges as if from thin air. It can seem, from popular media, that data science can perform miracles. The truth is data science is more about guiding conversations toward truths rather than producing truths. Beyond a somewhat mythical view, I don't know if I knew that much about the actual practice of data science. Mean – Standard Deviation – Probability – Naïve-Bayes – F-Statistic: all of these terms were Greek to me. (And, I actually KNOW Ancient Greek). This innocence served me well as I entered the ADS program with an open mind.

## *APPLIED DATA SCIENCE AT SYRACUSE UNIVERSITY*

The Applied Data Science program at Syracuse University's School of Information Studies equips students with the opportunity to develop analytic, technical and managerial skills in order to contribute measurable impacts in a highly competitive job market. The focus of the program is encapsulated in the word "Applied": students are taught to collect, manage, analyse and develop insights that can be communicated clearly to a broad audience of interested parties. Through courses such as Data Administration Concepts & Database Management (IST 659), Data Analytics (IST 707), Scripting for Data Analysis (IST 652) and Advanced Database Management (IST 769), I learned basic data structure (from byte to bit); how relational databases are built using SQL, Excel and Microsoft Access, meeting the business needs of organizations; how data mining techniques can be used to solve classification problems using R Studio; and, scripting in Python to explore a dataset to gain insights; using visualization to explore data and to communicate findings to a broad spectrum of audiences.

The Applied Data Science Program has seven learning objectives which were exemplified by the applications in this portfolio:

1. Describe a broad overview of the major practice areas in 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 relevant professionals in their organization.
7. Synthesize the ethical dimensions of data science practice.

# *PROJECT 1 – IST 659 DATA ADMINISTRATION CONCEPTS & DATABASE MANAGEMENT*

## Course Description

IST 659 is an introductory course to database management systems. This course examines data structures, file organizations, concepts, and principles of database management systems (DBMS) as well as data analysis, database design, data modeling, database management, and database implementation. More specifically, it introduces hierarchical, network, and relational data models; entity-relationship modeling; basics of Structured Query Language (SQL); data normalization; and database design. Using Microsoft's Access and SQL Server DBMSs as implementation vehicles, this course provides hands-on experience in database design and implementation through assignments, lab exercises, and course projects. This course also introduces advanced database concepts such as transaction management and concurrency control, distributed databases, multitier client/server architectures, web-based database applications, data warehousing, and NoSQL.

## Learning Objectives:

After taking this course, the students will be able to:

- Describe fundamental data and database concepts
- Explain and use the database development lifecycle
- Create databases and database objects using popular database management system products
- Solve problems by constructing database queries using Structured Query Language (SQL)
- Design databases using data modeling and data normalization techniques
- Develop insights into future data management tool and technique trends
- Recommend and justify strategies for managing data security, privacy, audit/control, fraud detection, backup and recovery
- Critique the effectiveness of DBMS in computer information systems

## Deliverable(s) – Custom Built SQL Data Base

Working under the tutelage of Dr. Gregory Block, I developed a database for a hypothetical real estate developer to manage: various communities; homes for sale within those communities; agents; buyers; all culminating in sale transactions. The goal of my project was to sort out some potentially complicated relationships (homes that are part of a community but can be sold individually, entities who could own or buy a house, and agents who could represent buyers or sellers).

With assumed client consultations, the project developed business rules which helped guide the first a conceptual model of the database, followed by a more refined logical/normalized model. Afterwards, I created tables and populated data using SQL Server Management Studio. Custom views and stored procedures were created as well as connection to Microsoft Access and Excel to produce reports to answer data questions, such as inventory of homes the developer/client has available, sales that have occurred and a detail of sale transactions.

## Learning Outcomes & Reflection

| | | | | | | |
|---|---|---|---|---|---|---|
| ● | ● | ○ | ● | ● | ● | ○ |
| Describe a broad overview of the major practice areas in data science. | Collect and organize data. | Identify patterns in data via visualization, statistical analysis, and data mining. | Develop alternative strategies based on the data. | Develop a plan of action to implement the business decisions derived from the analyses. | Demonstrate communication skills regarding data and its analysis for | Synthesize the ethical dimensions of data science practice. |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | relevant professionals in their organization. | |

The process of developing business rules, and each stage of the logical model to normalization, proved essential for creating a seamless data insertion process. I utilized Access and Excel ODBC connections to create my reports relying heavily on Views I created. A deep understanding of Excel/Access as well as the operations of a real estate firm helped me to complete these tasks.

*Project 2– IST 652*
*Scripting for Data Analysis*

## Course Description

Scripting for the data science pipeline. Acquiring, accessing, and transforming data in the forms of structured, semistructured, and unstructured data. The goal of this class is to teach students the tools and skills of scripting needed to solve problems of accessing and preparing data in a variety of formats and situations, sometimes known as data wrangling. The scripting will provide the skills needed to form data science pipelines, from acquiring and cleaning data to accessing data and transforming data for analysis or visualization.

## Learning Objectives:

Upon successful completion of this course, the student will be able to:

- Write scripts to access and amass data from fields in structured data, access fields in semistructured data, and define and find patterns of data in unstructured data; prepare and transform data to produce data summaries, lists, and networks;
- Analyze and solve data access problems for the three types of data and to find and deploy appropriate software packages that can be integrated into the problem solution; and
- Frame real-world data questions and show how they can be answered from data.

## Deliverable(s) – Final Project Proposal, Project Code in Python, Project Report & Presentation

### Project Description

Working in groups, your final project you will demonstrate your ability to write Python scripts to access and amass data from fields in one or more of the three types of data studied in the course and to prepare and use data to produce data summaries, lists, and other structures. Pick a topic of investigation and the data that you will use, ideally from more than one source. The topic could focus on one main data set but also have supporting data. Your topic may focus on a single target topic or person, combinations of them, a comparison of more than one target topic and person, or comparisons over time. The data may come from any source: those that you have found online, collected from social media, or obtained through other means.

Pick several possible methods of analysis in order to give some initial idea of what analysis you will try. This analysis will be to answer the types of questions that you have worked on for the homework assignments. Since we are not focused on visualization, the results of your analysis can be reported as structured tables with a unit of analysis and collected, summarized, or computed values for those units.

### Team Project with Katie Haugh and Sandy Spicer

Under the direction of Dr. Landowski, my team decided to focus on Squirrels. The Squirrel Census, an Atlanta-based organization with a mission to count and acknowledge the Eastern gray squirrel, set out to expand their mission to the Big Apple. In 2018, an 11-day squirrel census was conducted in Central Park in New York City. The organization engaged 323 volunteers who surveyed the 843-acre park for squirrels. The 323-volunteer squirrel sighters were assigned shifts in the morning and afternoon to count squirrels. Each was assigned 1 hectare (100-by-100-meter blocks) per shift to conduct their observations for 20-25 minutes. Each countable hectare in the park – 350 in total – was counted twice (in the morning and late afternoon). The census team assigned a random number to each hectare, which signified the order to survey each hectare. The numbers were randomly assigned to remove any selective bias.  The National Oceanic Atmospheric Administration's National Weather Service station near Belvedere Castle collected the weather data from hectare 17-D. The total time sighters spent counting squirrels was nearly 368.7 hours. This project presents the survey results.

The project entailed ingesting two data sets in CSV format: Squirrels and Hectares. Squirrels contained data on location, age, color, and other characteristics of 3,023 squirrels found in Central Park. Hectares data included location, weather, and environmental information on the 350 hectares of Central Park during the census. Our team performed preprocessing steps, determining the meaning of each variable, finding variables that contained categorical data points aggregated into one cell, which would need to be normalized. Also, we made decisions on data inputs like "unknown," which we determined to be different from "NA". Hectares had a wide range of weather observations which we relegated to discrete easy to understand buckets. Also, we dealt with formatting issues on temperature to remove the "°" and "F" characters. Our data set was nearly entirely categorical data except for temperature.

We answered these data questions:

- What are the primary visual characteristics of the squirrels? What color combinations were most popular? Were squirrels of similar markings found in a particular area of the park?
- Is there a section of the park that tends to have more squirrels? What are the characteristics of those hectares? Were other animals present?
- Is there a relationship between the number of squirrels seen and the time of day? Does the weather at the time have any influence?
- What verbal noises were most popular? What behaviors are most common? How do these distribute around the park?

## Learning Outcomes & Reflection

| | | | | | | |
|---|---|---|---|---|---|---|
| ○ | ● | ● | ○ | ● | ● | ○ |
| Describe a broad overview of the major practice areas in data science. | Collect and organize data. | Identify patterns in data via visualization, statistical analysis, and data mining. | Develop alternative strategies based on the data. | Develop a plan of action to implement the business decisions derived from the analyses. | Demonstrate communication skills regarding data and its analysis for relevant professionals in their organization. | Synthesize the ethical dimensions of data science practice. |

Even a low stakes project like Squirrels could be a candidate for robust data scientific exploration and analysis. This project was an opportunity to learn data processing techniques in Pandas, make decisions on how to handle sticky data problems, and organize our data to be able to create visualizations and a compelling story. My team desired materials that could have just as slick an approach as if we were pitching a Class A office tower leasing assignment in New York City…just, it's about squirrels! Visual storytelling is a crucial ability that a Data Scientist can offer.

*Project 3– IST 707*
*Data Analytics*

## Course Description

Introduction to data mining techniques, familiarity with particular real-world applications, challenges involved in these applications, and future directions of the field. This course will introduce popular data mining methods for extracting knowledge from data. The principles and theories of data mining methods will be discussed and will be related to the issues in applying data mining to problems. Students will also acquire hands-on experience using state-of-the-art software to develop data mining solutions to scientific and business problems. The focus of this course is on understanding of data and how to formulate data mining tasks to solve problems using the data.

The topics of the course will include the key tasks of data mining, including data preparation, concept description, association rule mining, classification, clustering, evaluation and analysis. Through the exploration of the concepts and techniques of data mining and practical exercises, students will develop skills that can be applied to business, science or other organizational problems.

## Learning Objectives:

After taking this course, the students will be able to:

- Document, analyze, and translate data mining needs into technical designs and solutions.
- Apply data mining concepts, algorithms, and evaluation methods to real-world problems.
- Employ data storytelling and dive into the data, find useful patterns, and articulate what patterns have been found, how they are found, and why they are valuable and trustworthy.

## Deliverable(s) – Final Project Classification Project Code & Report

The objective of the project is to use the main skills taught in this class to solve a real data mining problem. Students can choose to work individually or pair up with another student.

- Checkpoint 1: project idea proposal and presentation: Your idea proposal should include an overview of the data mining problem, the data set you will use and its availability, and your proposed data mining approach.
- Checkpoint 2: project progress presentation: Show preliminary results and major challenges.
- Checkpoint 3: Final project report: The final project report should describe the data mining problem, its significance and broader impact, the data mining approaches, results, and interpretation of the discovered patterns.

### 2019 Boston Marathon Dataset

Under the direction of Dr. Block, I focused my project on the 2019 Boston Marathon Results dataset. I pulled the data from a CSV file posted on Boston Athletic Association (which manages the event) website www.baa.org ; I wanted to pull raw data from a primary source as opposed to from GitHub, to have the full experience of a data set which might require extensive cleaning. After some experimentation with clustering that did not yield results (discussed with the professor at Checkpoint 1), I decided to make use of the commonplace classifiers inherent in running data (Gender, Age, Country). In researching methods of analyzing the data, I found some data scientists had created "Splits" data by arithmetic means. Dividing the finish time by 26, produces an average AKA Pace. Pace could then be used to calculate 5K, 10K and Half-Marathon splits. One problem: that data is not actually provided in the data set. Dr. Block and I decided that the results gained by such a method are problematic. The stakes of the results of a race are low, but transferred to another dataset, it's the pathway that a data scientist can allow bias to take root in analysis. We decided that the data scientists role is only to seek insights available from the data itself. "Noodling" is not allowed.

That said, I sought an separate BAA source (Qualification Times by Age/Gender Group) and applied a process to create a new variable "Achieved Qualifying Time" with respect to Gender and Age group. In other words, to qualify for the race a runner would have needed to reach a particular time, indexed by their particular grouping. I created my classification models to predict whether a runner could actually achieve their Qualifying Time. Using factors such as Age Quartile, Gender, Region, I used Decision Tree ("DT"), Support Vector Machine ("SVM") and Random Forest ("RF") and achieved generally 80% Accuracy rates with moderately good Kappa. An initial test with Naïve-Bayes yielded inferior results and was not included. Retuning my parameters yielded slightly better results.

## Learning Outcomes & Reflection

| | | | | | | |
|---|---|---|---|---|---|---|
| ○ | ● | ● | ○ | ● | ● | ● |
| Describe a broad overview of the major practice areas in data science. | Collect and organize data. | Identify patterns in data via visualization, statistical analysis, and data mining. | Develop alternative strategies based on the data. | Develop a plan of action to implement the business decisions derived from the analyses. | Demonstrate communication skills regarding data and its analysis for relevant professionals in their organization. | Synthesize the ethical dimensions of data science practice. |

This exercise gave me a wealth of hands-on R keyboard experience. From solving the problem of accessing data, to translating a key continuous variable (finish time) from HH:MM:SS to a discrete numeral (number of seconds) to deciding to utilize domain knowledge that I have over competitive races, my data science skills grew leaps and bounds. Understanding the path from data collection to exploration to modeling to find an answer will often raise as many questions as are answered. A data scientist must be ready to adapt to the challenges that a particular dataset gives us. For instance, we rely on a binary variable such as "Gender", which in this case, is still received as a binary. There will come a day where gender spectrum will need to be respected and a binary variable may not be what we can count on. Data scientists must be prepared for this.

*Project 3– IST 769*
*Advanced Database Management*

## Course Description

An analysis of relational and nonrelational databases and their corresponding database management system architectures. Learn to build complex database objects to support a variety of needs from both the big data and traditional perspectives. Data systems performance, scalability, and security.

This course provides tour of relational, document, key-value, columnar, and streaming database systems through the lens of the CAP theorem. We will explore the strengths and weaknesses of various database systems in the relational, Hadoop, and noSQL spaces. Where possible you will experience these systems first-hand as to gain an understanding of how they can be used to address complex, big data challenges.

## Learning Objectives:

After taking this course, the students will be able to:

- Understand advanced issues with the relational database model, such as transactions, performance, and security, as to understand the need for other database models.
- Explain the CAP theorem and describe how any given database system's architecture fits within the CAP context.
- Compare different database models such as document, key-value, column-family, streaming, and relational.
- Identify the most suitable database systems for a specific application's data storage requirements.
- Evaluate relational, Hadoop, and noSQL database tooling as to understand their underlying similarities and necessary differences.

## Deliverable(s) – Weekly Coding Lab Assignments

Lab Homework assignments are technical activities which enforce asynchronous concepts through practice and are based on the demos in the course videos. Homework must be completed before the week's live session, where students must be prepared to discuss the outcomes of the assignment.

## Learning Outcomes & Reflection

| | | | | | | |
|---|---|---|---|---|---|---|
| ○ | ● | ○ | ● | ● | ● | ● |
| Describe a broad overview of the major practice areas in data science. | Collect and organize data. | Identify patterns in data via visualization, statistical analysis, and data mining. | Develop alternative strategies based on the data. | Develop a plan of action to implement the business decisions derived from the analyses. | Demonstrate communication skills regarding data and its analysis for relevant professionals in their organization. | Synthesize the ethical dimensions of data science practice. |

The lectures (both live and asynchronous) and particularly the lab work provided me the opportunity for me to deepen my understanding of SQL, going into advanced topics not covered in IST659, such as transactions, views, stored procedures, temporal tables and security measures. The coverage of this material reinforced that the data scientist must be a steward of an organization's data, its most precious resource, and become familiar with ways to ensure its security. Digital threats are more the norm now than ever before. Also, organizations need to be concerned with scale and useability. SQL may not always be the solution. Exploration of the CAP theorem, and exploration of ETL/NoSQL programs like Hadoop, MongoDB and others provided an invaluable view into ways to structure data to meet data needs that may not be satisfied by RDBMS.