Portfolio Milestone

FALL 2020

Matthew A Beck

GRADUATE STUDIES CANDIDATE | SYRACUSE MASTER OF SCIENCE IN APPLIED DATA SCIENCE

MBECK07@SYR.EDU

SUID: 549206035

Introduction

My name is Matthew Beck, and currently reside in Washington DC. I have studied data analysis since 2011 and have been practicing as a full-time data analyst since 2015. Thank you for taking the time to read this submission—I invite you at any point to contact me concerning my valuable experience studying for a Master's Degree in Applied Data Science at Syracuse University, and how the program has benefit me as a lifelong learner and in my current work as a manager in business intelligence.

In my initial application to the Applied Data Science program at Syracuse, I highlighted my excitement to develop my skills in an environment that blended existing theory with practical application. I'm proud to say that as I reflect on my time in the program, I feel appropriately equipped to tackle any challenges that come my way as I progress further in the field of Data Science. I'm equally excited to share my journey with you in the context of this overview. To that end, I've chosen to structure this discussion around the core competencies (program learning goals) from which this degree program is centered, which I've summarized below:

Core Competency	Course	Projects
Describe a broad overview of the major practice areas of data science	• IST 687 – Applied Data Science	• All
Collect and organize data	• IST 659 - Data Admin Concepts & Database Management	• Metrics Engine
Identify patterns in data via visualization, statistical analysis, and data mining	• MAR 653 – Marketing Analytics	• Customer Segmentation
Develop alternative strategies based on the data	• MBC 638 – Data Analysis and Decision Making •	• Process Improvement Project – Reducing Survey Creation Time
Develop a plan of action to implement business decisions derived from analyses	• IST 718 – Big Data Analytics	• Hops and Hatha Image Recognition
Demonstrate communication skills regarding data and its analysis for relevant professionals in their organization	• IST 772 – Quantitative Reasoning in Data Science	• State of Vaccination Rates in California School Districts Report
Synthesize the ethical dimensions of data science practice (e.g., privacy).	• IST 659 – Data Admin Concepts & Database Management	• Metrics Engine, Personal notes from live lectures

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

From the very first lecture in my introductory data science course (IST-687), data science has been characterized as an interdisciplinary area – a successful data scientist must possess a strong background in the "**collection, preparation, analysis, visualization, management, and preservation of large collections of information**" (Salz, IST 687 Week 1 Lecture). The projects I've chosen to highlight from my experience at Syracuse all possess six elements to some degree, and I will outline the connection within their relevant contexts.

Collect and Organize Data

### Metrics Engine – IST 659 Data Admin Concepts & Database Management
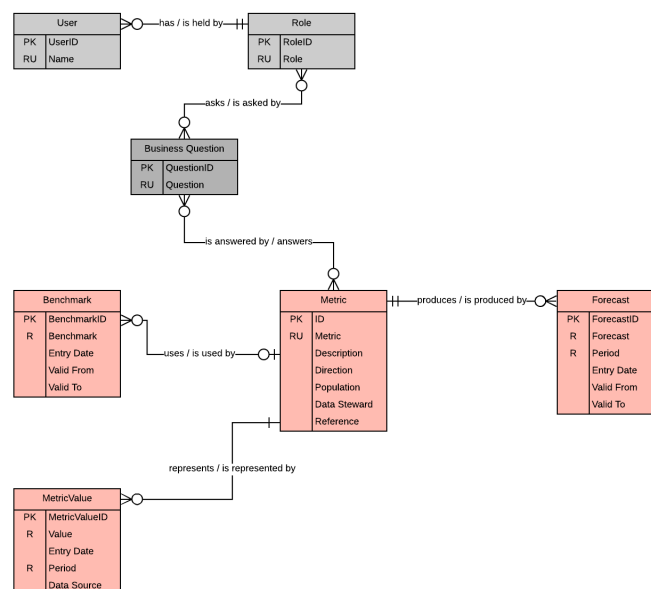
In my current work, I report to a company made up of a complex hierarchical structure with many business leaders overseeing one or many departments. Historically, success metrics and their associated visual representation have varied significantly from one area of the business to the other. The goal of this project was to link these metrics to the vital questions of the business in an extensible framework that could be easily accessed and updated by question askers and those responsible for answering. By simplifying both the collection and display of these metrics and their associated metadata, I can help leaders in the business focus on driving action instead of sifting through dense analytic reports.

This project directly tied to the collection, management, and preservation of data for the long term, and was immensely valuable to my work in Business Intelligence. One of the main issues I face is finding a computationally light way to capture and display vital data points to the operations of the business. This course taught me how to implement a relational database solution from the ground-up, and I currently maintain this database using an RDS instance in Amazon Web Services. To be successful here, I was required to apply the theories of data

management discussed in class, and implement a logical model derived from higher-level discussion with the business. I currently utilize the lessons I learned here in my current work and intend to leverage solutions like this for future stakeholders desiring the same positive outcomes achieved in this effort.

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

Customer Segmentation – MAR 653 Marketing Analytics
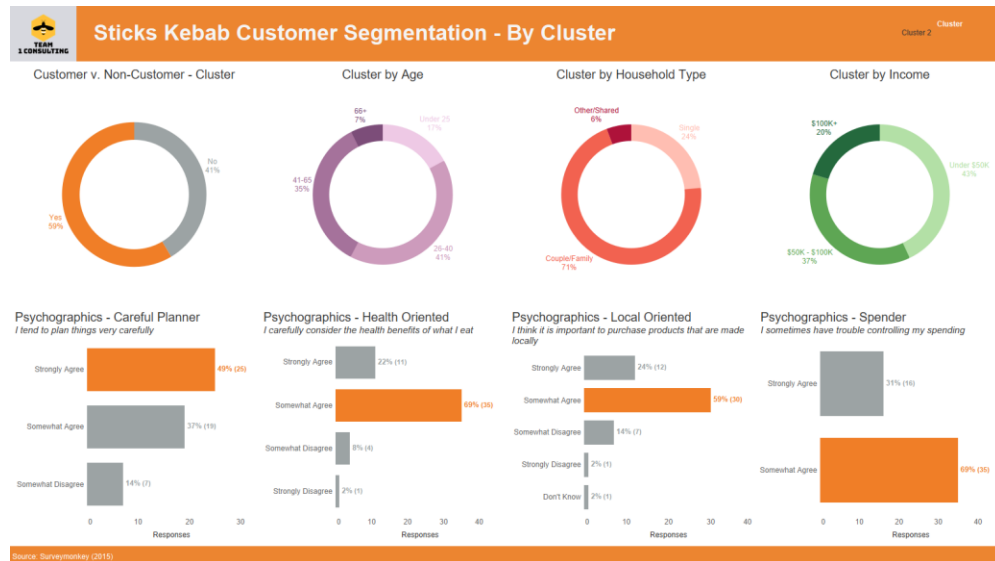
In this exercise, the task at hand was to take regional survey data collected by a local restaurant group and make recommendations on where they should open future locations. This work also required the use of K-means clustering to derive the optimal number of customer segments, the chart below informing the team's decision on how many segments to derive.
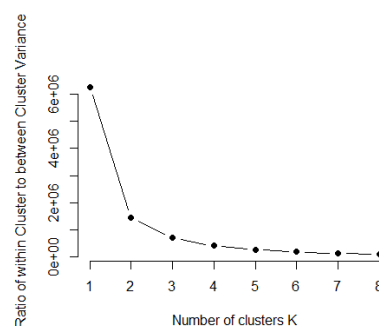
My specific contribution to this project included the **preparation, analysis,** and **visualization** of the supplied survey data to help the client make the optimal decision. I found this exercise to be a positive contribution to my existing work in data science, as I took the skills learned in this exercise to help make recommendations to my own company on their segmentation strategy. I enjoyed using Tableau as a medium both for data preparation and for visualization, and R for heavier computation around clustering.

Develop alternative strategies based on the data

Process Improvement Project – Reducing Survey Creation Time

Define Oct 7 - Oct 20	Measure Oct 21 - Nov 3	Analyze Nov 4 - Nov 17	Improve Nov 18 - Dec 1	Control Dec 2 - Dec 19

Define

Our existing process for survey development and distribution has been identified as an issue with several potential points of failure, including:
- Lack of measurement on distribution efficacy or quality
- Decentralized documentation around survey/stakeholder responsibilities
- Lack of standard distribution timeline
- Manual quality assurance without a documented standard practice

Cross-team collaboration in this process may be a factor in time required

CURRENT PROCESS SQL:

2

Measure

I measured the existing process by collecting exported data from:
- Ticketed requests for surveys in JIRA (project management software)
- Survey team feedback on successful and defective surveys

Survey Creation Distribution

Distribution does not look 'normal'

Survey Creation Duration (days)

Avg. Survey Creation Duration (2018)

Spike in June!

Analyze

I analyzed this process using the following approaches:

Measures of Central Tendency	
Mean:	18.9
Median:	14.9
Mode:	13.9
Measures of Variation	
Range:	60.4
Standard deviation:	13.6
Variance:	184.2

High variance indicates an inconsistent process!

This analysis helped me prioritize the areas where I could provide feedback on where to approve. Proposed changes needed to include:
- A goal set significantly lower than the existing mean
- Access to real-time data to help prioritize tickets during times of high volume (e.g. June)
- Brainstorm ways to reduce burden on the survey team

Improve

The team sought to improve this process by instituting the following ideas:
- Set a team goal for 2 week (14 day) turnaround on all survey requests
- Use data from JIRA to measure which tickets were at risk for failing to meet the goal to help prioritize work.
- Requiring teams to include more information on the ticket itself, reducing time the survey team spent manually entering the information.

Since implementation, initial signs indicate the changes may be starting to work!

Box & Whisker Plot

Drop in range, outliers

Old Process   New Process

As more data become available, a new SQL will be calculated.

Control

I ensured that the changes can improve the process over time by instituting a standard framework that determines whether the process is in control:

Process Control Chart (Individual)

I — Xbar — UCL — LCL

Process Control Chart (Moving Range)

mR chart — mR bar — URL

Next Steps:
Calculate new SQL after 6 months of new process
Review Process Control Charts monthly
Calculate new measures of variation and central tendency after 6 month mark.

This exercise offered another excellent opportunity to apply the theory in class to current problems in my role. My company creates and distributes customer experience surveys on a regular cadence. These surveys are the backbone of our company's ability to interpret the quality of our product offerings. The team in question had complained about several points of failure, which in turn could create a risk to the company's ability to measure not only customer satisfaction, but also to measure the impact of executive strategic decisions in response. Resolving these issues will decrease the required billable hours per survey and reduce the overall cost of the survey process-both big wins for the company.

To develop alternative strategies for the team, I **collected, prepared, analyzed,** and **visualized** raw data from the business, and made recommendations based on what I discovered. The initial findings were positive, and control measures were put in place to ensure that the recommended changes had the desired effect. The DMAIC framework taught in this course is a valuable framing for improving process, and I've found continued value in applying it with other clients.

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

Hops and Hatha – IST 718 Big Data Analytics
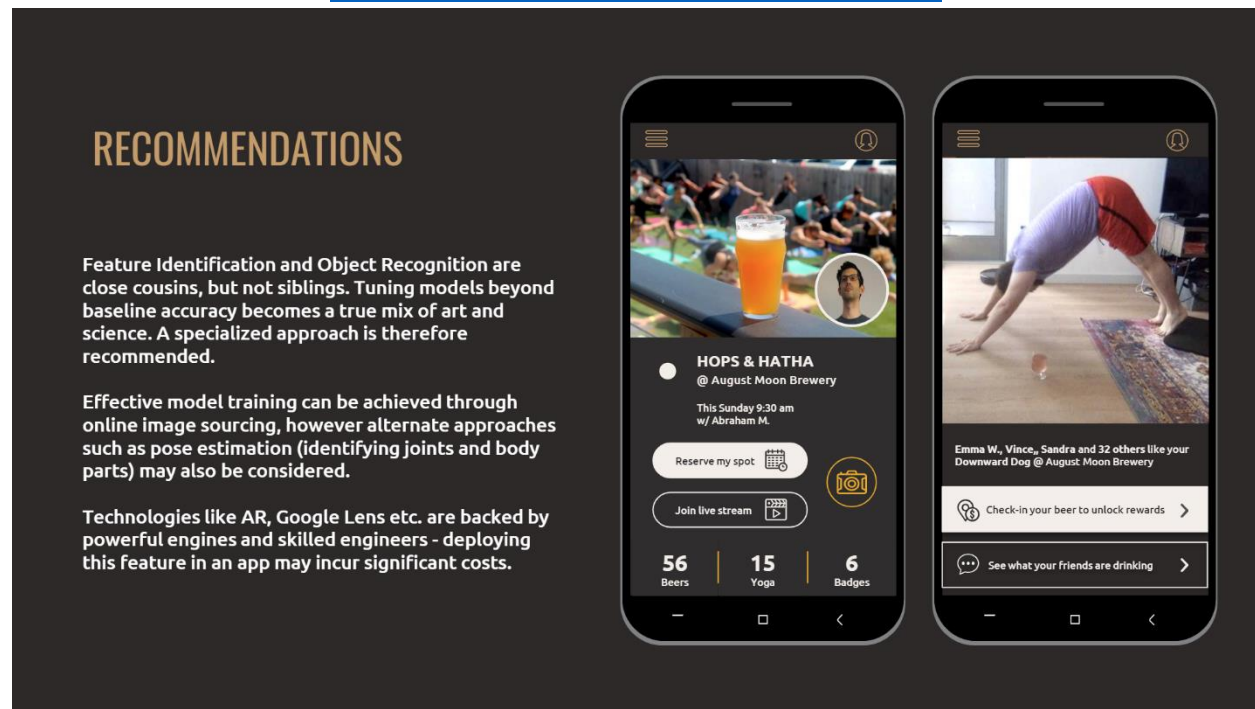
In this project, I collaborated with a team on behalf of a fictitious client seeking to utilize Machine Learning for desired features in an upcoming app. My personal contribution to this project was to develop the business case and associated recommendations based on the developed model's output—leaning heavily on my **analysis** of the data.

Our team ultimately developed a model that would automate identification of a given yoga pose, and a model that can recognize beer, whether bottled, canned, or glass. Based on the relatively strong performance of the models separately vs. together, I made the recommendation to use both models separately rather than seek to implement them together, ensuring that the maximum accuracy would be used in the final product. My work in this project was a useful lesson in letting the data guide the ultimate approach, rather than let the initial specifications dictate the finding.

Demonstrate communication skills regarding data and its analysis for relevant professionals

State of Vaccination Rates in California –IST 772 Quantitative Reasoning in Data Science
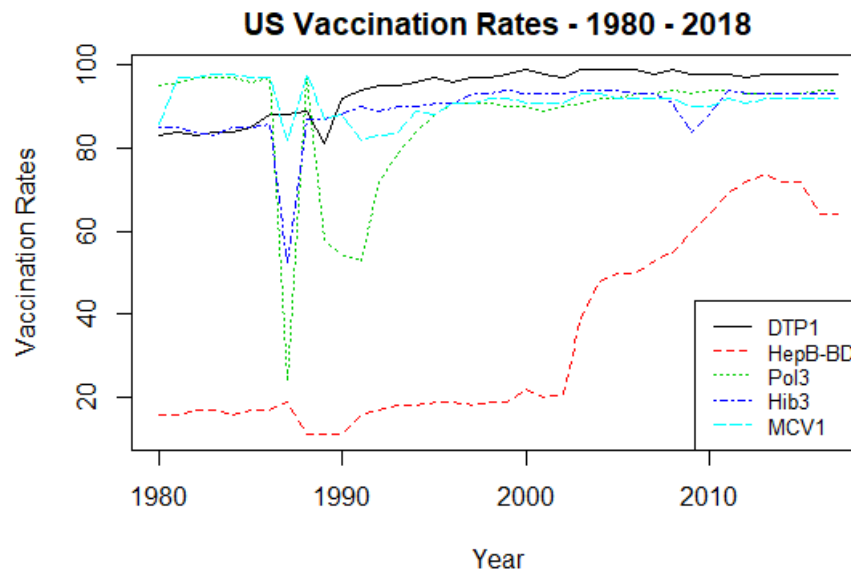
For this project, I **prepared**, **analyzed**, and **visualized** three datasets related to vaccination rates in the United States and within a selection of school districts in California. I applied varying methods of statistical analysis to examine the relationship between reporting completion, reported vaccination completion rates, and belief-based exemption rates across California school districts using an array of available demographic variables. This analysis indicated potential relationships between:

* Vaccination reporting completion and the ratio of students per school in the district

* Reported vaccination rates and both in-district free meal eligibility and student to school ratio (positive relationship)

* Belief-based exemptions and children under the poverty level in the district (positive relationship), in-district free meal eligibility and student to school ratio (negative relationship)

I ultimately recommended my intended audience (a state legislator) allocate financial assistance to school districts in rural areas with high proportions of individuals (especially children) who are under the poverty level.

In order to be successful here, I needed to combine multiple methods of analysis (Bayesian, Frequentist) to answer multiple research questions in a way that a less data-savvy audience would understand. I found it to be effective preparation for other reporting needs in my position and found the class itself well-suited to presenting data to wider audiences.

Synthesize the ethical dimensions of data science practice

Metrics Engine – Revisited

Executive - Primary audience for business KPIs, reviews them with managers to gauge department performance

Manager - Relies on KPIs to measure their department's performance, reviews with executives, analysts, and staff.

Analyst - Helps identify trends and produces reporting for managers to use in conversations with executives.

Staff - Individuals who generate the actual business output measured by KPIs.

Exploring a different aspect of the Metrics Engine project – In order to successfully implement the data structure I had designed, I first had to understand how to translate business rules into role-based access. Reflecting on this exercise in **managing** the data and associated access brought me to review in-class discussions I'd taken part in concerning the ethical use of data. From that discussion I wrote the following:

Practitioners in data science have an obligation to accurately and faithfully represent the data obtained, in every step of the process. Simply having access to data does not mean it should be used, especially if its use creates negative impact.

This perspective has stayed with me through my course in this program, and absolutely guides the decisions I make today. Simply put, data scientists have an obligation to analyze and represent their data both in terms of how the world currently is, and where the world could be. Part of my mission in my field is to exercise wisdom in ensuring that data is not used to perpetuate existing problems, especially when it comes to analysis of critical areas like medicine, poverty, law enforcement, and climate change.

Final Thoughts

Reflecting on my experience as a student in this program, I'm grateful for the opportunity to hone my craft and apply this knowledge to life moving forward. Not only has this program helped improve my data literacy, but it has provided all the necessary tools to collect, prepare, analyze, visualize, manage, and preserve data with great confidence and expanded understanding. I look forward to applying these methods to my remaining work in the semester, and in the opportunities to come.