**Portfolio   
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Syracuse University

School of Informational Studies

Keeley Ables  
M.S. in Applied Data Science Student  
khables@syr.edu

**Table of Contents**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Introduction ……………………………………………………………………………………………………………………... 3-4

Project 1 Reflection …………………………………………………………………………………………………………. 4-12

Project 2 Reflection ………………………………………………………………………………………………………... 12-24

Project 3 Reflection ……………………………………………………………………………………………………….. 24-37

Influential Readings ………………………………………………………………………………………………………… 37-38

Conclusion …………………………………………………………………………………………………………………………. 38

References ………………………………………………………………………………………………………………………… 39

Appendix ……………………………………………………………………………………………………………………… 40-89

**PORTFOLIO**

**INTRODUCTION**

In this work, I seek to reflect on and demonstrate my learned proficiency of the expected outcomes specific to the Applied Data Science master’s program, of which I am a candidate. My objective being that the reader may understand how the project deliverables, along with their associated courses, have contributed to this learned proficiency. In addition, I strive to persuade the reader of the influence of texts on my acquired competency concerning the learning objectives of the program. Furthermore, upon completion of this reflection, the reader should be able to gauge my growth in these competencies in light of my increased prospective professional capacity. In exploring these particulars, I aim to portray my growth as a future data analyst/scientist in the field of education.

To preface my analysis aligning project deliverables to the Applied Data Science program objectives, I present my initial baseline of understanding concerning data analysis. I consider this preface relevant for the purposes of assessing growth and learning. It is difficult for me to explain how I have learned specified objectives and how the completion of projects has contributed to my “ultimate present abilities” if initial abilities at the start of my engagement with this program are not understood. In short, if the starting point is not known, growth is not appreciated. Consequently, I begin my reflection with this topic.

At the start of this program, I was a third-grade teacher. The entirety of my exposure to a professional career consisted of five-and-a-half years of teaching third and fourth grades. As a public-school teacher educating students from low-income (Title 1) and typically low-performing schools, I was *extremely* familiar with handling, extracting insights, and generating instructional practices, from data. After all, educational best-practices require a teacher to refer to student data to drive all instructional decisions. This is especially important in schools that typically experience limited student achievement. Despite this focus on student data, and despite my undergraduate education (B.S. in Alternative Medicine) being fairly math and science intensive, I had never taken a complete statistics course. Furthermore, in the field of elementary education, attempts were certainly made at quantitative thinking. However, few elementary educators were trained appropriately for such analysis. As a result, the realm of public elementary education, as I experienced it, recognized the necessity for data-driven initiatives, but was limited by the scope of academic preparation had by most educators in this area. Consequently, in keeping with the culture in my field, I put an inappropriate amount of emphasis on correlative relationships between various potential predicting factors and the dependent variable that was student performance. I lacked the skills to quantitatively look further at interesting relationships. Additionally, I lacked the knowledge to appropriately consider data in context. I, like many other teachers, sat in many staff meetings anxiously as I tried to formulate an explanation as to why my student scores had increased one week, then decreased the next. None of us considered the possibility of natural variability, or the need to plot the data on a control chart to attempt to differentiate signals from noise. However, this time in education was a fitting transition to my time in the Applied Data Science program because there are multiple high-level parallels between student performance improvement and the contribution of data science concepts to business problem solutions.

As discussed earlier, this reflection of my work in the Applied Data Science program discusses how various projects I have completed, along with their associated courses, align to the expected outcomes of my studies. For ease of reference, I include the program’s expected outcomes/objectives here, as found in the *Project Portfolio Milestone Requirement Detailed Description and Procedures* from Syracuse University specified for the Winter 2022 semester:

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 managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization.
7. Synthesize the ethical dimensions of data science practice (e.g., privacy).

For further ease of reference, expected outcomes may be referred to throughout this reflection as Objective One, Two, etc. as they align with the numbered order displayed above.

**PROJECT 1: MBC 638 The DMAIC Project**

Given my background, MBC 638 Data Analysis and Decision Making was the perfect choice as one of the first classes beginning my journey into data science. This class communicated not only the importance of data collection, but how to consider it in context. It introduced me to some foundational statistical concepts which would support efforts to explore data in search of initial insights and discern which analyses could most appropriately answer business questions. With the skills taught in the Data Analysis and Decision-Making course, I successfully completed the (Define Measure Analyze Improve Control) DMAIC project. This DMAIC project illustrated an alignment with the expected outcomes for the Applied Data Science program outlined at the start of this reflection, and consequently allowed me to demonstrate my abilities regarding these outcomes. Specifically, I opted to use DMAIC to improve my BMI (body mass index) by lowering it to a healthier range. In consideration of Objective One concerning the ability to “describe a broad overview of the major practice areas in data science,” this project exhibited my ability to use data analysis to indicate and implement improvements regarding a process. This operations/ process improvement aspect may easily be considered a major practice area of data science, as the use of its concepts translate to increasing efficacy in said processes/operations. I detail specific improvements and the analysis driving this improvement in later paragraphs to prevent redundancy. The reader will find that my discussion of the project and my ability to improve my process through analysis of the data, will be sufficient evidence of skills needed for Objective One.

The collection of data and ability to access said data using the appropriate technology were competencies added to my skill set as the course developed. Additionally, this project enabled the opportunity for demonstration of these newly acquired abilities that aligned closely with Objective Two of the expected outcomes. As referenced at the start of this reflection, Objective Two requires the student to “collect and organize data.” I concede that the scale of data collection was miniscule in relation to the likely intentions of Objective Two. However, one must consider that this was my first exposure to data collection for the purpose of quantitative analysis. Although frequent student data collection was a staple of my past profession, there was no expectation of statistical analysis performance after this assemblage. Therefore, the DMAIC project introduced me to the importance of data consistency and format for the purposes of quantitative analysis using a given technology. To elaborate, the histogram, control charts, correlation, and regression analysis utilized in this project would all be impossible if the numerical data inconsistently contained commas, or erratic units of measure. Additionally, the data used was small and simple enough that Excel was the applicable technology used. Likely Excel was the introductory tool for the Data Science program so that analysis could be mastered on this simple but still in-demand technology. In these ways, the DMAIC project addressed Objective Two of the expected outcomes.

I further contend the DMAIC project’s support of the program’s expected outcomes by discussing its alignment with Objective Three. Although the DMAIC process does not engage the entire life cycle of data science, there is indeed overlap. To begin, if using the Specify Observe Analyze Recommend (SOAR) model as a guide through the data science process, the Specify and the Define steps of the SOAR and DMAIC processes respectively can be likened to one another. Both steps communicate the business problem and its impact. In this project, my business problem was that my BMI rested in an unhealthy range. The business impact was that I was at a higher risk for cardiovascular and metabolic diseases, which could cost myself and my family $120,000 to millions of dollars over the course of my life. The Observe step in the SOAR model and the Measure step in DMAIC also share similarities, as they both are inclusive of exploratory data analysis. Exploratory data analysis in my project took the form of a histogram displaying a distribution of my BMI data points, as well as a control chart. These visualizations from the original project are included below in figures 1-2.

**Figure 1. Histogram**

Although this histogram is not a distribution visualization in the purest sense, it does convey that the distribution of my data possibly had a slight right skew, which likely affected the median of the data. This is especially true when considering the small size of the data.

**Figure 2. Control Chart**

This ImR control chart was used to evaluate my initial improvement process prior to analysis. Although it seemed my process was in control, the chart raised the question about days 2-5 and days 7-10. Did these days align with certain practices (independent variables) in my process?

It seemed the primary difference between the DMAIC and SOAR is the latter concludes with a recommendation and former executes an implementation then works to maintain improvements. Because regression from the Analysis step in the process suggested that the intake of calories shared the only statistically significant relationship with the dependent variable BMI, I implemented a healthy limit of 1,500 calories per day in the improvement phase. As a result, I surpassed my goal for lowering my BMI by 1.5kg/m^2. The result was such an improvement that, continuing the project would require a new control chart with lower maximum and minimum limits. It is the above explanation that convinces me that process improvement (using DMAIC) is closely related to the life cycle of data science, and furthermore, that this project along with the course skills it demonstrates is responsible for my ability to improve a process and monitor continuous improvement/maintenance.

**Figure 3. High-Level Storyboard Overview of BMI Project Using DMAIC Process**

A picture containing graphical user interface

Description automatically generated

Although the above discussion alludes to objective three of the expected outcomes, as visualizations (histogram and control chart) are displayed to provide insights during exploratory analysis, I directly address how my DMAIC project focusing on BMI data supports this objective here. During the Analysis step of the DMAIC project, I searched for correlative relationships in Excel to gain insights into possible correlations between the independent factors that I had identified and the dependent factor BMI. I did this to see if correlative relationships would provide any clues to potential influential factors on BMI. Of course, there was no expectation of the correlation analysis to reveal causal factors of increasing/decreasing BMI, as it is well known among statisticians that correlation does not equal causation. The following return from the analysis is found in Figure 3 below.

**Figure 3 Correlation Analysis**

**Graphical user interface, application, table, Excel

Description automatically generated**

The correlation analysis conducted in Excel, found on the right side of Figure 3, displayed the correlation values between each of the independent variables highlighted and the dependent variable, BMI. My criteria for a “strong” correlative relationship was > |0.7|. Using this criteria, Early Bedtime, Carb Intake, and Calorie Intake were eligible for deeper investigation.

Figure 3 displays the independent factors Exercise, Early Wake Up, Early Bedtime, Carb Intake, and Calorie Intake along with their associated correlative relationships with the dependent variable BMI. Notice that Early Bedtime, Carb Intake, and Calorie Intake were the variables with the highest correlation. This provided the clue that statistically significant factors may be found among these three. To “gain actionable insight” on how to improve my BMI level, I further investigated the relationship between my independent and dependent factors by running regression analysis. Although there were only three independent factors that, using my criteria of ≥ |0.7|, were considered to have strong relationships, I did use all the factors in the regression model to make sure I was not excluding other statistically significant variables. Figure 4 below displays the return from this analysis.

**Figure 4. Regression Analysis With all Independent Variables**

Table

Description automatically generated

From this analysis, I noted that the only statistically significant variable boasting a P-value < 0.05, was the Calories variable. With this new insight, I visualized the linear relationship between Calorie intake and BMI. Below, Figure 5 displays this visualized relationship and the prediction it provides.

**Figure 5. Graph of Linear Regression using BMI and Calorie Intake**

The visualized analysis in Figure 5 uses the relationship between the explanatory and dependent variable to predict the BMI if calories are increased. Specifically, the model predicts that for the consumption of every 1,000 calories, my BMI should increase by 0.05kg/m^2. This is true with a constant BMI of 39.851 kg/m^2. From the predictive regression model, I can interpret the R^2 value to mean that the model can explain approximately 50-51% of the data. With a now, somewhat more experienced eye, I concede that I would probably have used a different type of regression than linear. Experience in this program inclusive of executed coursework and projects, has taught me that there may be a better fit for this model. However, at this point in my journey through this program, the repertoire of analyses within my skill set, was limited. Furthermore, the return from the correlation and regression analyses did address objective three. As discussed, I was able to use the return from these analyses, along with the graph in Figure 5 to glean patterns and gain actionable insights. As mentioned earlier, my process was improved by the realization that calorie intake was the statistically significant variable influencing my BMI. These findings gleaned from relationships between x and y variables in the data exhibited my experience with Objective Three via the DMAIC project.

Not only did patterns and relationships found in visualization and statistical analysis drive improvements in the DMAIC project, they also contradicted my hypothesis concerning factors influential on BMI. This drove my process to focus on a healthy reduction of calories, as opposed to the reduced carbohydrate diet that I had implemented at the start of data collection. In a professional setting, I would recommend to stakeholders that deeper analysis be conducted. This lower calorie initiative was certainly effective for me; however, my sample of data was quite small. In addition, one person is hardly representative of the entire population. Furthermore, in hindsight with more experience analyzing data, I realize the potential for bias if one person is collecting data describing his/herself. Nevertheless, my analysis prompted the change from a low-carbohydrate diet to a lower-calorie diet. The result surpassed my goal for a decrease in BMI range. Below, in Figure 6, the new process map reflective of data analysis, implementing healthy lower calorie meals, is exhibited below.

**Figure 6. Improved Process Map Including Lower Calorie Meals**

Diagram

Description automatically generated

Although exercise was not shown to have a statistically significant influence on BMI, I did not consider it a healthy practice to remain sedentary for ten weeks for the purpose of isolating the Calorie variable. Consequently, during the Improve phase of DMAIC, I did not collect data on the Exercise variable but did include it for my own health. Exercise was often a leisurely walk.

**T**he results of this process improvement were the revised BMI data, displayed in Figure 7, against the previous upper and lower control limits as well as the previous mean of the data.

**Figure 7. Control Chart with New Data Against Old Control Limits**

Note that the new BMI data exhibits multiple points that fall well below the lower control limits, indicating the signal that the improved process is likely working.

Clearly, the improved process was effective to such a degree as to require a new control chart with new upper and lower control limits. That very chart is presented below in Figure 8.

**Figure 8. New BMI Data Plotted on New Control Chart with Updated Upper and Lower Limits**

Had I continued the Improvement phase, by the look of data points from days 10-12, a new control chart would likely have been necessary.

Because an alternative initiative influenced by data analysis was executed, as evidenced by Figures 6-8, I was able to gain experience implementing Objective Four and Five through the completion of the DMAIC project. These objectives required the student to pivot strategies based on the data and to formulate recommendations/implementations driven by analyses. Consequently, this project provided opportunity for me to practice both expected outcomes.

Objective Six was also addressed in the DMAIC project, evidenced by the presentation of my journey and findings through the DMAIC steps responding to my business problem. The project presentation was designed and should be comprehensible to listeners and stakeholders at various levels within the organization. Although, the presentation may best be understood by those with some exposure to applied statistics, the visualizations used were easily comprehensible and the analyses used were explicitly discussed. Concerning the last objective of the expected outcomes, I did implement fairness and transparency in my slides describing the project data, analysis, and findings. At this point, I had very little experience with different types of bias. However, I was aware of the personal nature and potential for bias considering it was my own BMI data. Furthermore, I was quite aware of the sensitivity of the data and inappropriateness of the project objective and disclosure of the data, had it belonged to anyone besides myself. In this way, the DMAIC project also addressed objective seven.

If I proceed in the field of education as a data analyst/scientist (these types of positions are becoming increasingly popular in public education), I will easily be able to extrapolate this skill set and target it to student growth/proficiency, along with other business problems found in the field of education. For example, this DMAIC model could certainly work to improve the infamously sluggish process of recommending students for services needed to support success. The process includes recommendation, testing, then implementing the needed supports to the identified students. This is a process that often begins at the start of an academic year, but the student may or may not receive these needed supports by the end of the same school year. The DMAIC process improvement framework would likely make this procedure much more efficient. In conclusion, this project laid down the foundation for my skill set regarding data science objectives, and this skill set could be easily translated to supporting student achievement with data analytics/data science within the field of education.

**Project 2: IST 687 Introduction to Data Science Project**

IST 687 was my first exposure to using the data science life cycle to address a business problem. In this course, I worked through each step in the process to complete a project with a team of my classmates taking the same course. Specifically, I collaborated with L. Co, C. McConnell, A. Schambach, M. Stockhaus, and G. Uhrig. The project was focused on discovering factors influential on school grade (derived from student standardized testing performance data) and using relationships to make predictions about future performance. This analysis utilized student elementary school data from the state of Florida. Because the data was used at my suggestion, and because of my professional experience in this field, I acted as the subject matter expert. Apart from this role, we all participated and collaborated equally, without explicitly assigning roles. As I participated in more projects as the program progressed, I became wise as to the benefits of assigning roles with associated specific duties. However, thankfully in this scenario, my team was knowledgeable, helpful, and hardworking. So, the lack of assigned roles for this project did not detract from achieving the objective. Incidentally, data wrangling, cleaning, visualizations, and analysis was conducted in R.

Like the DMAIC project, the Introduction to Data Science project continued to address the Applied Data Science objectives specified in the program. Here I assert that the project supported learning Objective One, as it provided the opportunity to experience the generalizations of the major practice areas in data science. Specifically, this project enabled me to implement data mining, statistical analysis, data visualization, and machine learning. As evidenced by the discussion below as well as the actual project materials provided in the appendix section of this reflection, we utilized data mining to find patterns in the data through various analyses. Because I elaborate on data mining, statistical analysis, data visualization, and machine learning as I proceed to explain how this project led me to gain efficiency in each learning objective, I omit an explanation concerning this project’s relationship with such major practices of data science in this portion of the reflection to prevent repetition. Regarding the Introduction to Data Science project’s alignment to program objectives, the reader will see detailed evidence of my experience with the above listed data science practice areas relevant to this project as this discussion progresses.

The Introduction to Data Science project allowed me experience with data collection, storage, cleaning, and wrangling, with appropriate technology, as specified in Objective Two. Additionally, as promised above, I offer this discussion as evidence of the project’s alignment with Objective One, as well. This data preparation is a needed step in data mining, one of the major data science practice areas I referenced in the discussion associated with Objective One. It is apparent that the team did not collect and enter student performance data from state standardized assessments in the ten-week time frame allowed by the course. However, data was collected from the Florida Department of Education website. The actual link to the data can be found in the appendix section of this reflection, in the reference portion of “Project 2.” Comprising one data set was an attribute communicating grades that the state assigned each school according to their students’ performance on different components of the Florida Standardized Assessment (FSA). This data set contains such data from all public and charter schools in the state for a particular year. Also included in said data were the various attributes Income, school-associated economically disadvantaged data, percent of student population who were minorities, regions, number of students in a school, the school’s status as a public vs. charter school, and the school’s status as a Title 1 vs. non-Title 1 school. A Title 1 school is one whose student body is composed of large amounts of low-income students. Consequently, these schools receive extra government funding for necessities such as free/reduced lunches and other supplies to support its students. These attributes would later be used as potential explanatory variables for student performance. Because only one year’s worth of this data was included in a data set, it was necessary to integrate all the data sets describing the years 2008-2019 together to form one large data frame. Additionally, zip codes, along with socioeconomic data associated with these zip codes, were joined with the combined performance data to create geolocation visualizations for analysis. Therefore, data collection consisted of joining the above data from disparate sources in Excel, then reading the culminating result into R. The combined table was stored, and accessed in R. This discussion justifies my assertion that this project aligned with and therefore provided opportunity to practice Objective One and Two, as it allowed me to practice the data preparation stage of data mining and data visualization, and it allowed me to collect and organize data.

In support of the program learning Objectives One and Three, the Introduction to Data Science project provided the occasion to use the SOAR model to work through the data science life cycle. The alignment to the two mentioned objectives lies in the close relationship between the process of data mining to discover patterns in the data and the phases of the SOAR model. Furthermore, the project enabled me, in collaboration with my team, to gather “actionable insights” concerning predictors of student performance using the data science life cycle, which incorporates visualizations and predictive models in gathering these insights. We began by articulating the project background, scope, and business problem, which basically communicated that we were testing our newly acquired skills in R to work with a challenging data set for visualization and analysis. We specified the breadth of the project in our scope, which expressed that the focus of analysis was on school grade data from Florida from the years 2008-2019, for the purpose of discerning which, if any, factors could be predictors of student performance. This understanding of possible predictive factors would provide insight into optimizing educational funds allotted for improving student performance, as specified in the business problem. These introductory steps align nicely with the Specify phase in the SOAR model, guiding us through the data science life cycle. Additionally, they instituted the objective for the data mining, visualizations, statistical analyses, and machine learning that would be conducted over the course of this project.

After wrangling and cleaning the data in preparation for the next phase of the SOAR model, the Observe phase took the form of control charts, distribution visualizations, correlation analysis, bar charts, an animated gif, and geolocation visualizations. This phase of the life cycle used exploratory analysis to provide initial insights into how the data may answer the business problem. Prior to exploratory analysis, we used criteria to cluster schools into the categories Stable, Improving, Declining, Inconclusive, and Fluctuating. Figure 6 below details the criteria for each category.

**Figure 6. Criteria for Categories of Schools Using their Student FSA Data**

Table

Description automatically generated

**Schools Aggregated by Categories**

Table

Description automatically generated

Note the small number of improving schools as opposed to the much greater number of declining schools.

The categories were added to the data set as an attribute and used for the following exploratory analyses. The control charts seen in Figure 7 show the movement of the school performance data from 2008-2019 by assigned category.

**Figure 7. Control Charts for School Performance Data by Category**

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

True to its category name, the “Improving Schools” increase student performance, most notably from 2014-2019. Interestingly, in 2014 Florida switched standardized testing from the Florida Comprehensive Assessment Test (FCAT) to Florida Standards Assessment (FSA). The switch in assessments allowed the state to assess the newly implemented standards which were closely aligned to Common Core standards. The increased student performance for this cluster of schools occurred that same year.

Notice the consistent variation in performance from 2008-2019 for the schools categorized as “Stable.” This variation is what may be expected of schools with either no initiatives, or ineffectual initiatives implemented.

Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated

Some event must have happened in 2012 (before I started teaching), as the significant decline in student performance that dips below the lower control limit in 2015-2016 is certainly a signal. The partial recovery of student performance from 2017-2019 is likely some sort of intervention initiative implemented by the district and/or state in response to the sharp decline from the previous years.

There are seemingly multiple signals in this data set. The improvement in performance from 2015-2019 could have been caused by the Florida assessment switch in 2014. This increase would certainly warrant further investigation. Additionally, the sharp decline from 2012-2014 likely was a signal of some happening occurring during this time frame.

These exploratory visualizations provided insight into the possibility of some happening that may have occurred from 2012 to 2014. The cluster of schools categorized as “Stable” was the only category that consistently performed through the alleged occurrence in this time frame. Investigation into these signals may have provided clues as to other possible variables influential to student performance in Florida. We continued to use histograms and boxplots (distribution visualizations) to make observations and glean initial insights. Several examples of these distribution visualizations are displayed in Figure 8 and 9 below.

**Figure 8. Distributions of Mean Incomes by Region**

Chart, scatter chart

Description automatically generated

Note the means of all five regions rest at approximately $50,000/year. Also notice the regions with extremely long tails in their distributions. The outliers are likely skewing the mean towards higher income values. The bulk of the data across the five regions rests in a lower income range.

To get a true sense of the center of the data concerning income by region, the median was found and can be seen in Figure 9 below.

**Figure 9. Median Income by Region**

Text

Description automatically generated

**Figure 10. Distributions of Economically Disadvantaged Percentages by Region**

Chart, box and whisker chart

Description automatically generated

With the mean income across regions concentrated in the range of roughly $40,000- $47,000/year, there would likely be a high percentage of economically disadvantaged students across regions.

In this Observe portion of the data science life cycle my group and I included many other visualizations focusing on various attributes we were using in analysis as potential explanatory variables on school grade. These attributes were listed earlier in the discussion. The complete project inclusive of all visualizations utilized to gain initial insights about these attributes during exploratory analysis, can be found in the appendix section of this report. However, here I include a sufficient sample to support my assertion of this project’s alignment to Objectives One and Three of the expected outcomes. Specifically, we used the data science life cycle, which includes using data mining and visualizations to identify patterns in the data. Because the Observe stage in the life cycle focuses on the exploratory analysis of the data, the given sample of visualizations provide insight into the independent variables we analyzed. To elaborate I highlight the initial categorization performed with the schools. We sorted them into categories such as “Improving”, “Declining”, etc. This process revealed that the number of improving schools (3) was miniscule in relation to the number of declining schools (124). This simple aggregation warned that Florida had an academic performance problem within public and charter elementary schools. We then focused on the income distributions by region. These yielded the understanding that across all regions of Florida, household incomes are considerably less than the nation-wide average of $68,703 in 2019, according to the United States Census Bureau (2019). Furthermore, our distributions of economically disadvantaged percentages by region revealed that all regions suffered a high percentage of economically disadvantaged students resting in the range between 75-100% across all regions. This is true despite the fact that each region had long tails that likely skewed the mean percentages to be lower than the true center of the data. To summarize insights gained thus far, we found that Florida elementary students were struggling academically. Very few schools demonstrated consistent academic improvement. We also saw that many of Florida’s students are economically disadvantaged. At this point in the data science life cycle, we saw a hint that there may be a correlative relationship between student performance and economically disadvantaged students. As correlation is not equivalent to causation, we had no evidence thus far that the variable describing economically disadvantaged students was a predictor of school grade. However, insights gained here did direct us to consider this variable as a possible factor in our investigation.

The Analyze step in the data science process contributed to further understanding of the educational situation in Florida. Below Figure 11 exhibits multivariate regression conducted with the independent variables listed at the start of the discussion of this project, and the dependent variable school grade.

**Figure 11. Multivariate Regression Using the Relationship with Independent Factors on Dependent Factor Student Performance**

Text

Description automatically generated

From this predictive model, we found the following as statistically significant variables with P value < 0.05:

* School’s status as Charter school Y/N
* Income
* Percent of economically disadvantaged students
* Percent of students who were minorities
* Region 4
* Region 5

Note, however, in the analysis above the adjusted R^2 value is 0.4688, meaning that these factors only explained about 47% of the data. We thought it necessary to delve deeper in analysis and hone our focus. As a result, we used regression analysis to predict student performance within the achievement clusters to which we assigned Florida elementary schools. These clusters were different from the improving, fluctuating, declining, stable and inconclusive categories discussed earlier. Here we simply categorized schools as Excellent, Declining, or Other. Attempting to focus analysis for more accurate results, we used regression to predict student performance within Excellent and Declining schools, as these were the clusters of most importance. Below in Figure 12 and 13, are the regression analysis for both classes.

**Figure 12. Regression Model to Predict Student Performance for Excellent Schools**

Text

Description automatically generated

The adjusted R^2 value resting at 0.4678 experienced almost no change. However, this analysis was able to home in on three statistically significant variables that are predictors of student performance. These are percent of student population economically disadvantaged, Region 2, and status as a Title 1 school.

**Figure 13. Regression Model to Predict Student Performance for Declining Schools**

Text

Description automatically generated

In this model, the only statistically significant variable is percent of students economically disadvantaged. However, note the adjusted R^2 value is only 0.3611. This communicates that only an approximate 36% of the regression line is explained by the data. This value suggested that linear regression was not the best fit for the data since all the adjusted R^2 values for all the regression models were low. However, these models consistently directed our attention to the variable that described the percent of students economically disadvantaged. In search of better accuracy and confirmation of our findings from the regression analyses, we conducted decision tree and random forest analyses seen in Figures 14 and 15 below.

**Figure 14. Machine Learning Decision Tree Analysis to Predict Most Influential Variable and Student Performance.**

Diagram

Description automatically generated

Note once again percent of economically disadvantaged students appears as the most important variable.

**Figure 15. Random Forest (Importance of Variables)**

Chart, scatter chart

Description automatically generated

Once again, these visualizations which are products of predictive analysis, are but samples of the full range of analysis performed on the data to address the business question. However, they serve to illustrate how analyses directed us to the conclusion that percent of economically disadvantaged students was the variable that was not only consistently statistically significant, but also consistently predicted that student performance would improve as the percent of economically disadvantaged students declined. The above discussion is evidence that this Introduction to Data Science project directly supported the referenced Objectives One and Three. This project required me to practice Objective One, as it utilized some of the “broad overview of the major practice areas in data science” such as data mining, visualization, and machine learning. The skills I gained and utilized in this project aligned with the skills referenced in Objective Three, as it involved “identifying patterns in data visualization, statistical analysis, and data mining.”

Although the pattern from the analyses suggests that the variable describing economically disadvantaged students holds the greatest statistically significant influence over the variable school grade,, the accuracy/level of fitness of the regression line from the earlier models must be discussed. As noted in the earlier discussion, with R^2 values as low as 0.3611, it was clear to my team and myself that the relationship between the independent and dependent variables were not necessarily linear. In fact, as I reflect on this analysis, I concede that instead of performing linear regression within each cluster of schools, I would have performed multivariate logistic regression. This is because the categorization of schools into one of the assigned clusters would have been a prediction concerning school performance and the regression line would have probably been a more accurate fit for the data. However, at this stage of my learning journey, logistic regression was outside of my newly acquired skill set. As a result of the lack of a linear relationship between the variables, we focused on other models, such as the machine learning decision tree displayed in Figure 14. The team continued to build other machine learning models, such as a Kernel Support Vector Machine (KSVM), Support Vector Machine (SVM), Naïve Bayes model, and Random Forest. As a result of the inaccuracy seen in the regression models, these machine learning models were used to predict the category of school performance specified earlier. The two labels in question were “Declining,” “Excellent,” and “Other.” Figure 16 below displays the model predictions and their accuracy evaluations. This pivot into different types of analyses driven by reflection on the evaluation of the regression models exemplifies my experience with Objective Four via this project.

**Figure 16. Machine Learning Models Predicting School Performance Categories**

**Text

Description automatically generated with medium confidence**

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Given the results of the analyses discussed above, it may stand to reason the recommendation addressing the business problem included the redistribution of state educational funds to provide additional instructional support for schools with large concentrations of economically disadvantaged students. The instructional support should take the form of funds for resources, but also additional instructional support staff that could aid teachers with working in differentiated small groups with students, targeted to their proficiency needs. This would be more efficient than current practices inclusive of district abandonment of the previous year’s costly initiatives to implement new ones, with the assumption that the previous initiative was ineffective because scores decreased relative to the previous year. Currently, in my professional experience, little instructional differentiation at the district level is implemented for schools whose student populations suffer higher percentages of economically disadvantaged students. Although Title 1 schools do receive additional funding, this often is focused on providing free breakfast and lunch to all its students, as well as supporting clothing needs, healthcare needs for students, after school tutoring for English Language Learners (ELL), and the like. These funds are not necessarily assigned for additional instructional interventions/supports or the additional professional development for teachers, needed to make the accelerated gains with students to approach grade level proficiency in the classroom. These actionable insights were directly derived from the visualizations, predictive models, and the data mining process which was encompassed in Objectives One and Three in the expected outcomes. The above discussion is evidence of how insights gained from the project’s analyses drove the action plan specified here, which was the requirement for Objective Five. Therefore, the Introduction to Data Science project provided the opportunity for me gain experience with said objective.

Additionally, this endeavor was presented to an audience of my peers, along with my instructor. Because I knew my audience and I knew it was necessary to display knowledge of statistical and machine learning concepts learned in the program, my team and I presented the project accordingly. Knowing my audience also meant awareness that few of them would be familiar with the professional setting in the realm of public education. Therefore, the presentation contained explanations concentrated on domain-specific educational concepts. Less of an explanation was provided for statistical and machine learning concepts, as my instructor along with the rest of my peers would share familiarity with such ideas. Therefore, this presentation would likely be suitable for audiences composed of statisticians and possibly IT professionals and programmers, as referenced in Objective Six, depending on the scope of their background and training. However, had I presented this project to school board members or a super intendent for a school district I would have reversed my strategy. I would have concentrated my explanations on statistical concepts and only focused on presenting the most comprehensible visualizations which require no proficiency in statistics. In this scenario, the situation could be considered comparable to communicating with managers, as referred to in Objective Six. Therefore, I successfully succeeded in discerning how to communicate with various roles appropriately and comprehensibly in a professional capacity. It seems that this was the aim of the sixth learning outcome.

Finally, the Introduction to Data Science project supported Objective Seven by protecting the privacy of instructional staff and students, explicitly limiting our recommendations to Florida public and charter schools, accepting the conclusions of data analysis despite the disparity between the results and my subjective experience with student performance, and submission to ethical transparency of data analysis.

Data collection protected the privacy of students and teachers using aggregation at the school level and grade level. It may seem intuitive that the disclosure of individual student data is illegal. Therefore, to avoid this unethical, illegal practice, student data is aggregated at school and grade level on the Florida Department of Education website. Teacher privacy is protected as well. In the field of public education, student test scores are utilized to hold teachers accountable, much the same way that sales staff in a business would be held accountable by their sales. Student performance accounts for half of a teacher’s evaluative status in various Florida school districts. If a teacher’s student data fails to meet expectations, that teacher will often be moved to a grade level that is not tested. In extreme cases the teacher may even experience a non-renewal of his/her annual contract. This means his/her contract will not be renewed the following academic school year by the same school. The teacher would then be free to seek a contract with another school, which often is quite difficult, as concealing a non-renewed status from administrators in the same school district is quite difficult. Consequently, district/state-wide disclosure of a teacher’s student data carries the potential for embarrassment and shame. In some districts, the union advocates for a clause in the teacher contract that prevents disclosure of student data at the class level in association with the class’s teacher outside the school for this reason. For these privacy reasons the lowest level of disclosure concerning student data is at the grade level for each school. This is because the grade level would contain multiple classes of student data, and it would be quite difficult to associate data with a particular student or teacher.

To further ensure the ethics and accuracy indicated in Objective Seven, conclusions and recommendations about exploratory/predictive/machine learning data analysis in this project were made specific only to Florida. To elaborate, conclusions and recommendations were focused only for the purposes of the Florida Department of Education. Our team did not presume that the influence carried by the economically disadvantaged status for students would extrapolate to rest of the United States (US) student population. As a result, Florida state educators were the only educators to whom we would report conclusions, despite how intuitive it may be that this variable could influence student performance across the country. Indeed, it would reveal bias on the part of our team, especially myself because of my experience in the field, because we conducted no analysis on how well Florida students would represent the general US student population.

My team and I continued to advocate for the ethics and accuracy of the analyses by allowing the data to drive conclusions rather than my subjective experience with student performance. In my experience, it is commonly understood among teachers that Title 1 schools are expected to perform lower than non-Title 1 schools. Of course, this conception is contested by some district and school-based administrators who encourage teachers to implement best-practices during instruction, instead of excusing themselves from extra work using this misconception. Many teachers, however, hold this perception as fact. The entirety of my personal teaching experience has been in struggling, Title 1 schools. I have seen these schools improve significantly and perform well. However, in my experience, teachers and administrators worked many extra hours relative to instructional and administrative staff at non-Title 1 schools, to obtain these significant results. Consequently, when the analysis of our data did not find a significant relationship between Title 1 status and student performance, regardless of my surprise, ethically it was necessary to adhere to the analysis results and make recommendations accordingly.

Finally, my team and I observed appropriate standards of transparency. As evidenced by our project report included in the index of this reflection, we included:

* The source of the data we used
* How it was joined, cleaned, and wrangled
* The exploratory analysis inclusive of its process, visualizations, and observations
* The models inclusive of their processes, visualizations, evaluations, conclusions
* Explanations for how one analysis drove another.

Suffice to say, the Introduction to Data Science Project supported Objective Seven of the expected outcomes, as it ethically managed data and the process included in the data science life cycle.

This project has obvious implications should I further my data science skill set professionally in the field of public education. I referenced the inefficient use of allotted funds on instructional initiatives with little to no appropriate quantitative analysis concerning factors influencing student performance. The budgeting of educational funds could be reformed. The age-old question that faces teachers at the start of every school year, asking “Should I focus on student growth, or student proficiency?”, could be answered. All educational resources (staff, textbooks, etc.) could be optimized. The data science life cycle including data mining, machine learning, visualizations, etc. translates well to asking and answering questions facing schools and districts, state education departments, and even educational data at a national level.

**Project 3: IST 718 the SOAR Project**

Although I have referenced the SOAR model in earlier discussions, IST 718 Big Data Analytics was my first introduction to this concept. This model is synonymous with the data science life cycle. Despite the apparent overlap between my two previous projects and this model, the course Big Data Analytics delved deeper into the data science life cycle using significantly larger data sets.

As one could guess, I have aptly named the SOAR project such because my team and I followed the SOAR model in its execution. My team consisted of D. Grey, D. Piston, and myself. The project’s purpose was to build a recommender system for a figurative movie streaming service such as Netflix. To do this, we used past user ratings data to create various predictive models of users’ preferences. The data was provided by Netflix on the Kaggle website below.

<https://www.kaggle.com/laowingkin/netflix-movie-recommendation/data?select=qualifying.txt>

The user data was a large collection of 480,000 randomly selected users from 1998-2005. The totality of the data equaled 100,000,000 rows. Data was read into Google Colab using Python as the language that cleaned/wrangled the data, analyzed it, and produced visualizations.

Whereas other projects seemed to enable the opportunity to practice the objectives for the Applied Data Science program, this SOAR project offered the occasion for mastery of said objectives. Because it was necessary for me to execute, then write about and present our project’s process, which included major practice areas of data science, I certainly gained the ability to “describe a broad overview” of them, as required by Objective One of the expected outcomes. To elaborate, I was able to master statistical analysis, machine learning, data mining, visualization, handling large amounts of data, and recommendations. Once again, as discussion progresses regarding the mastery of other objectives, I will highlight the mastery of Objective One where appropriate to prevent redundancy.

As SOAR is one model that implements the data science life cycle, and as the data science life cycle is closely related to the data mining process (a major practice of data science), it is necessary to explicitly discuss the project as it relates to each phase of the SOAR model. Indeed, this model displays the mastery of the objectives in expected outcomes of the Applied Data Science Program. The first phase of the SOAR model is Specification. In this step, the business problem outlines the objectives for the project. The Specify phase lays the foundation for the entire project. Therefore, it is necessary for the mastery of all Objectives, as it justifies their execution.

At the start of SOAR project, my team and I agreed that we would appreciate the opportunity to build a recommender system, as none of us had yet had that opportunity. We chose to build a model that would predict movie preference for a figurative movie streaming service, based on user ratings randomly chosen from the years 1998 to 2005. Initially, we searched through several movie user ratings data sets before finding one that contained the needed attributes User ID, Movie ratings, Movie ID, Movie Title, Year of Release (year movie was released), and Date (date movie was rated). In fact, when we did find this data via Kaggle, it was divided into four files allotted as the training data. A fifth file containing Movie ID, Movie Title, along with the Year of Release was provided for the purpose of being joined with the ratings data to associate the ratings with the actual movie titles. Also given was a probe file to be run after the model was trained on the training data. The seventh file was the test data set, which would evaluate the model. The data set containing movie titles could be joined to any of the training or testing data sets via the Movie ID attribute. In addition, the training data sets needed to be combined to form one large data set comprised of 100,000,000 rows. Each row described one movie rating instance. This large data set was converted to a pickle file for efficient storage. Subsequently, the newly consolidated training data, along with the file with the movie titles and testing data sets, had to be converted into a tabular format for analysis. Initially, all data sets were provided as Python dictionaries. They needed to be converted to data frames. This data wrangling exhibits the mastery of Objective Two which requires the collection and organization of data.

Once the data was wrangled and prepared for analysis, my team and I could then transition into the Observe phase of the SOAR model. This is where we would begin to implement some of those major practice areas of data science, specifically data visualization, that Objective One indicated would be necessary to describe. Additionally, this phase is where the project could begin to display mastery of Objective Three, as the Observe phase is where we would begin to gain insights by looking for patterns in the data. These patterns would be evidenced largely through data visualizations in this phase. Therefore, the Observe phase is where we began our exploratory analysis of the data. To elaborate, visualizations were used to communicate the frequency of movies in our repertoire by year of release, the distribution of movie ratings across all release years, the frequency of user ratings over time, the movie ratings frequency per movie count, and the movie ratings frequency per user count. Figure 17 below exhibits the frequency of movies by year of release. This would help us understand whether there was a uniform distribution concerning the year of release of the movies in our data, or if there was a time frame where the bulk of our data rested. We hoped this insight would allow us to gain some understanding of the distribution of the ratings across years of release.

**Figure 17. Count of Movies by Year of Release**

Chart, line chart

Description automatically generated

The line graph clearly displays the bulk of the movies in the data were released in the early 2000’s approximately in the range of years between 2000 and 2005. This may seem intuitive since the data was taken from the time frame between 1998 and 2005. In this scenario, it is certainly possible to have movies that predate these time frames, as evidenced by the line graph. However, movie content that would have been newly released during this range of years may have been more likely to enjoy higher popularity than older movies. Consequently, it is probable that more movies from the early 2000’s were included in the content repertoire to please customers. Furthermore, it may be the case that some sort of rights needed to be able to stream older movies, could possibly have been difficult to obtain.

Satisfied with this newly acquired insight, we proceeded to find the distribution of all movie ratings across all release years. Movie ratings were communicated by indicating an integer from one to five. A rating of one expressed the least favorable rating, whereas five expressed the most favorable rating. A count of these ratings was aggregated by movie rating and visualized as the simple bar chart seen in Figure 18 below.

**Figure 18. Distribution of Movie Ratings Across Release Year**

Chart, bar chart

Description automatically generated

The bar graph in Figure 18 displayed the tendency of users to rate movies at least somewhat favorably. This tendency suggested the possibility that users may favor newer releases, as the bulk of the movie content enjoyed a release date current to the time frame from which the data was taken, and users tended to assign somewhat favorable ratings to movie content.

To further support the above suggestion, the missing piece to this assertion would be the time period in which the bulk of the ratings were submitted. If the ratings data was spread uniformly across the range of years from which our data was taken, this would not support the seemingly emerging pattern that current movie content seemed to be most popular. Therefore, a time series analysis of movie ratings submitted between 2000 and 2005 was performed to visualize the spread of this data. The return from this analysis is visualized below in Figure 19.

**Figure 19. Frequency of Movie Ratings From 2000 to 2005**

Chart, line chart

Description automatically generated

It seemed there may be some support in the above visualization for the pattern suggested by the data that users may prefer to watch newly released content. The bulk of the ratings were submitted in 2005, the last year of the time period from which the data was taken. However, further investigation was certainly necessary.

To this end, we visualized the frequency of movie content ratings by movie count. The visualization of this analysis is displayed in Figure 20.

**Figure 20. Movie Ratings Frequency per Movie Count**

A picture containing application

Description automatically generated

This visualization suggests the bulk of ratings submitted are regarding a relatively few movies. The exception seems to be the obvious anomaly at the 10k value on the x-axis. Although, this anomaly was quite perplexing, its investigation did not necessarily support our objective. Therefore, we did not delve further into this question.

Figure 21 concludes our exploratory analysis of the data. It displays the frequency of user ratings per user count.

**Figure 21. Ratings Frequency per User Count**

Chart

Description automatically generated

Like Figure 20, Figure 21 suggests that a few users are responsible for the bulk of the user ratings. Also, like Figure 20, Figure 21 displays a similar obvious anomaly at the 200 value on the x-axis.

Through the Observation phase of the SOAR model, mastery of Objective One and Objective Three of the expected outcomes was evidenced by the execution of major practice areas of data science, including data visualization and steps included in the data mining process. Objective Three was evidenced by the pattern and insight gleaned from this data visualization and steps in the data mining process. Specifically, the patterns/insights gained from this particular exploratory analysis communicate the bulk of the movie content included in the data was released in the early 2000’s, between 2000 and 2005. Additionally, the bulk of the ratings were at least somewhat favorable. Furthermore, the bulk of the movie ratings were submitted in the 2004-2005 time period. These visualizations seem to suggest that users submitting user ratings seemed to favor newly released content. Visualizations from Figures 20 and 21 seem to suggest that possibly a relatively few users were rating a relatively few movies. However, as noted in the earlier discussion, further analysis was necessary to draw conclusions around models that would recommend user preferences.

Such further investigation would be performed in the Analysis phase of the SOAR model. This phase allowed me to demonstrate Objective One and Objective Three was similarly demonstrated via this SOAR project, as it portrayed patterns in the data through visualization, statistical analysis, and data mining. Furthermore, this project allowed me to display the flexibility to implement the data-driven pivots in approach to analysis referred to in Objective Four. The following discussion details the Analysis phase of my SOAR project, as well as its alignment to the referenced objectives.

Prior to implementation of various analyses, my team and I faced our first obstacle with the data. The entirety of the observations in the pickle file created in the wrangling stage totaled 100,000,000 rows. This made processing the data a serious challenge. Neither myself nor my team members owned a machine powerful enough to read in all the data in a reasonable time frame. To elaborate, 25% of the data was processed in no less than two and a half hours. In the face of this obstacle, we realized that it was necessary to down sample the data. At this point, we pivoted from our original, and somewhat preferable plan, to perform analyses utilizing the entire data set. Instead, we randomly sampled the data, using only 25% of it. We chose random sampling, as the only data we had surrounding users was their ID and rating information. We had no personal information for any sort of clustering, and therefore had no evidence that other sampling strategies would more sufficiently represent the population within our data set. Furthermore, due to the lack of information describing users, we had no data with which to use any sort of stratified sampling. This redirected strategy was demonstrated mastery of Objective Four of the expected outcomes. It could even be argued that some statistical knowledge would be necessary to discern what could be and could not be accomplished concerning sampling of the data. This statistical knowledge could be considered a modest degree of evidence of alignment to Objectives One and Three because knowledge of statistics is required in major practices in data science, as well as statistical analysis.

Our initial choice for a user recommendation model was to find the top-ten movies with the highest mean ratings, according to the user ratings data. This would give generic recommendations across users according to the most popular movie content. Figure 22 displays a bar graph displaying the top ten highest ranked movies across users.

To evaluate all the models utilized for the Analysis phase, the difference between the predicted values and the actual observations in each model, or the Root Mean Squared Error (RMSE), was used. We chose this evaluative measure because it is sensitive to outliers. Consequently, anomalous values in the data would carry heavy penalties in this evaluative score, making them quite noticeable. This heavy penalty was due to how the score is calculated. The Root Mean Squared Error is the square root of the squared difference of the sum of predicted and actual values in the model. Because the average is not taken in this calculation, outliers brazenly stand out. This heavy penalty for outliers was desirable to us because we wanted to be alerted to anomalies.

**Figure 22. Top Ten Movies Using Mean of Ratings**

Chart, bar chart

Description automatically generated

Consistent with patterns gleaned from exploratory analysis discussed earlier, note that much of the popular content above was current to the early 2000’s time period.

The RMSE is quite low at < 1, which suggests the model is accurate. However, the drawback of such a model is that it is a generic list of recommendations. This model does not differentiate according to user preference. This lack of differentiation would likely pose a problem when it comes to model maintenance. As noted earlier, much of the content exhibited in Figure 22 is composed of movies current to the time the ratings data was submitted. Although there is no apparent pattern concerning genres, content, or talent, the year of release recent to the data is common to most of the titles listed in the visualization. However, this model would require close supervision with constantly updated movies and ratings in order to be useful in the future. This is because new movie content is consistently released and without such updates, such a model could easily become obsolete. Additionally, this model may prove problematic if the ratings data contained bias. To elaborate, with only one small sample of the data analyzed, the effect of few users rating a small number of movies is unknown. It remains to be seen if this would be problematic to the accuracy of the highest mean model.

Attempting to improve the accuracy and usability of our model, next we created a weighted mean model. This is a model utilized by the free streaming service IMDb. It is quite similar to the highest mean model from Figure 22. It too returns the top-ranking movies, using the mean user ratings of movies. However, this model attributes less weight to data that appears to be anomalous with potential to skew the mean of the user ratings data. This weighting strategy works to preserve the quality of the data and accuracy of the model. Figure 23 illustrates the return of this weighted mean model.

**Figure 23.** **Top Ten Movies Using Weighted Mean of Ratings**

Chart

Description automatically generated

Upon examination of this weighted mean model, there are several points that became apparent. The first detail is that the RMSE may be slightly higher than the previous model. The increase might be considered miniscule, but it is notable that this model is slightly less accurate than the previous model. To elaborate, the error (distance between predicted and actual data) is slightly higher in this model, meaning that it is slightly less accurate. Furthermore, it is also notable that much of the movie content is consistent with the previous model, and much of it is current to when the ratings data was submitted. This seems to also support the pattern gleaned in the exploratory analysis suggesting that users seemingly preferred recently released content. Despite the high accuracy (RMSE score) of this model, I must highlight that the same inefficiency exists in this model as did in the highest mean model. That is that the weighted mean model is also a list of generic recommendations. There are no differentiated suggestions targeted for specific users. Additionally, because constant updates regarding new movie content, and new user ratings may prove model maintenance difficult for this model. Finally, like the highest mean model, the effect of few users rating few movies is unknown.

In the ongoing search for more accurate models that would also yield patterns and consequently insights into the data, next we executed a model that did differentiate according to customer preference concerning movie content. Our model recommended movies to a user based on content that the user had rated favorably in the past. The visualization for this model is seen in Figure 24 below.

**Figure 24. Ranking of Top 10 Recommended Movies for a User Based on Similarity**

Chart, bar chart

Description automatically generated

RMSE 1.3354

The glaring, undisputable takeaway when considering the bar graph visualizing the model from Figure 24, is the RMSE has, once again, increased. Although the RMSE may be regarded as evidence that the model, like the two previous, remains accurate because the error is still relatively low, notably, it has increased. This increase in distance between the predicted and actual data may (or may not) be due to the lack of accounting for user preference that enjoys a variety of genres and content. This model does not allow for user exploration of movies unique to past viewings. The model assumes that the user will only be interested in viewing movies similar to previous, favorably rated viewing choices.

In search of a model that would return differentiated recommendations, as well as boast greater accuracy, we built a Singular Vector Decomposition (SVD) model. This collaborative, machine learning technique gleaned patterns in movie content from favorably rated movies, and patterns in poorly rated movies, from past user ratings data. The model also gauged the relationship between the patterns and these ratings. It used these considerations of patterns to return its recommendations for each user. Because of the user differentiated nature of this model, a simple bar graph summarizing user preferences was not possible. Instead, Figure 25 below exhibits the return of the SVD model in Python.

**Figure 25. Singular Vector Decomposition Model**

Text

Description automatically generated

RMSE: 0.90+

Figure 25 displays the various RMSEs, Mean Absolute Errors (MAE), fit times, and test times. To further explain, because the model is fitted on previous user ratings data and makes recommendations (user preference predictions) at the user level, the return from this model includes many RMSE, MAE scores, as well values that communicate the time it took to fit each model (fit\_time), and the time it took to test each model (test\_time). The MAE returns the average of the differences between the predicted and actual values. Because it returns the mean of these differences, outlier data is less obvious than in the RMSE score. This explains why we chose the latter to evaluate the models.

In Figure 25, notice that the average RMSE score would be approximately 0.9. Although only slightly lower than the other models, this model does boast the lowest RMSE score. Consequently, this evidence suggests that it is the most accurate of the models thus far. This model is also the most differentiated modeled tailored to the user that had been run on the data, as it looked for relationships between patterns in favorably and unfavorably rated movies to predict user preference.

Finally, in the name of thorough evaluation and model comparison, we ran a cross-validation on various algorithms. The evaluations of each can be compared in Figure 26 below.

**Figure 26. Cross-Validated Comparison of Surprise Algorithms**

A picture containing table

Description automatically generated

Note the difference in the RMSE and MAE evaluative scores across all the models. Because the RMSE penalizes anomalous data more severely, it is intuitive that the RMSE scores would be higher than the MAE scores. Therefore, the RMSE scores give the appearance of a less accurate model.

Upon consideration of Figure 26, it is important to highlight that a random seed was utilized to run the models in Figure 26. This means that a different sample of data was used in the cross-validation visualization than the sample used to run the SVD algorithm from Figure 25. This sampling technique explains the inconsistent RMSE scores associated with the SVD models in Figures 25 and 26. Regardless of these inconsistent evaluative scores, the cross-validation visualization from Figure 26 supported the conclusion discussed after the consideration of the model in Figure 25. That is that the SVD model enjoys the lowest RMSE, and resultingly, the evidence suggests it is the most accurate model. Although the MAE was not chosen as the evaluative tool, we did observe that it was another bit of evidence that supported the seemingly superior accuracy of the SVD model.

The above discussion of the Analyze phase of this SOAR project yields evidence of mastery regarding Objectives One and Three of the expected outcomes. Data visualizations are demonstrated in Figures 22, 23, 24, and 26. Visualizations from Figures 22, 23, and 24 communicate the ten highest ranked movies according to various criteria specific to each respective model. These visualizations exhibited proficiency in the ability to discuss major practices in data science in that the data visualization, data mining, and statistical analysis encompassed within the Analysis phase of the SOAR project are such major practices, and in above discussions concerning these visualizations, I did thoroughly elaborate on them. Objective Three requires the data science student to “identify patterns” in data using visualization, statistical analysis, and data mining. Because of the overlap in Objectives One and Three, I address the alignment of the SOAR project with both of them in the following discussion. Gleaning patterns and insights is a fundamental reason for producing data visualizations, statistical analyses, and data mining. The patterns and insights gained included the following:

* Users seemed to prefer somewhat recent movie content.
* Movies enjoying the top ten status within each visualization in Figures 22, 23, and 24 appeared to have little in common in terms of genre, content, and talent.
* It appears that much of the data was collected and maintained accurately. There seemed to be few anomalies.
* All the models boasted low RMSE scores and therefore, appeared to be sufficiently accurate.
* Differentiated recommendations tailored at the user level, based on past viewings seemed to be less effective than the SVD recommendation model which considered relationships between high/low user ratings and the associated movie content.

*Pattern Observation: Users seemed to prefer somewhat recent movie content.*

This observation was discussed in the exploratory analysis of the SOAR project. However, it was also supported by the visualizations in the Analysis phase of this endeavor. This pattern was the most consistent factor that the top ranked movies within each respective model found in Figures 22, 23, and 24 have in common.

*Pattern Observation: Movies enjoying the top ten status within each visualization in Figures 22, 23, and 24 appeared to have little in common in terms of genre, content, and talent.*

To support this assertion, I draw examples from the models visualized in the figures in question. Figure 22 displays its top two ranked movie suggestions for users *as Family Guy: Freakin’ Sweet Collection* and *Lord of the Rings: The Fellowship of the Ring.* The content in these two works is quite different. *Family Guy* is television show belonging to the comedy genre. Lord of the Rings is the first movie of a trilogy, and it belongs to the fantasy genre. This difference in content exemplifies the significant variation in genre and subject matter between many of the movies displayed in the highest mean model from Figure 22. This lack of apparent similarity between the top ranked movies is sustained in the weighted mean model from Figure 23 and the similarities-based model in Figure 24. This pattern is consistent with the observation that the similarity-based recommender was seen to have the highest RMSE score, suggesting it may be the least accurate model. This may suggest that users tend to value a rich variety of different genres, subject matter and talent when choosing and rating movies favorably.

*Pattern Observation: It appears that much of the data was collected and maintained accurately. There seemed to be few anomalies.*

This assertion is derived from the weighted mean model seen in Figure 23, as well as the cross-validated comparison of algorithms seen in Figure 26. Recall that the weighted mean model was named such because it allotted less weight to anomalous, outlier data. In all other aspects, it is synonymous with the highest mean model. Yet when regarding not only the recommendations, but the RMSE scores from both models, they are quite similar. This means that there were likely few outliers skewing the means. Evidence of relatively few outliers exists also in Figure 26. Observe the consistently different scores between the RMSE and MAE scores across the various models included in the cross-validation matrix. Also, recall that the discussion earlier that explained that the primary difference between the RMSE and MAE evaluation measures was that the RMSE placed a heavy penalty on outliers due to the method of calculation. In contrast, the MAE score is the average difference between the predicted and actual value. Therefore, outlier penalties are not as significant. Despite these differences in calculation methods, Figure 26 exhibits little disparity between these two scores across all the models displayed. A small increase in all the RMSE scores is noted. However, if outliers were prevalent in the data, based on our premise that the RMSE carries a lofty penalty for anomalous data, there would likely be a wider discrepancy between the two evaluation scores.

*Pattern Observation: All the models boasted low RMSE scores and therefore, appeared to be sufficiently accurate*

Apart from the similarity-based recommender model with a RMSE score of 1.3354 (still relatively low), the other models from Figures 22, 23, and 25 boast a RMSE score of < 1.

*Pattern Observation: Differentiated recommendations tailored at the user level, based on past viewings seemed to be less effective than the SVD recommendation model which considered relationships between high/low user ratings and the associated movie content.*

In a previous pattern observation, I discussed that the data seems to reveal that users enjoy watching an assortment of different movie content. The similarity-based recommender model may be evidence that simply returning suggestions of movies similar to past favorably rated content could be less effective than the SVD model. Reasoning behind superior performance of the latter model could be that it explores the relationships between movie content that is favorably/unfavorably rated and its associated rating. The SVD model differentiates these suggestions at the user level.

The Recommendation phase of the SOAR model, which is derived from the patterns/insights previously specified, demonstrates my mastery of Objective Five of the expected outcomes. This objective requires me to devise an agenda to execute patterns and insights driven by analysis. To validate proficiency in this objective, I outline such and agenda in this discussion. My recommendations for this figurative movie streaming service include the following:

* Update movie content regularly and consistently. Ensure that a plethora of newly released content is available. Assure that newly released movies are adequately advertised.
* Advertise the benefits of rating movies viewed to users. Present an easy one-click rating popup at the conclusion of each movie viewing. Provide incentives for viewers to rate movie content.
* Analysis highlights the SVD model as the most accurate. However, stakeholders should evaluate the advantages/disadvantages of each model in terms of computational expense, time to run each model, etc. Additionally, assess the tradeoffs between choosing an appropriate model according to the needs of the business immediately, or evaluating each model on different samples of data to contribute more evidence that the SVD model is in fact the most accurate.
* Continue to update user-ratings data with current data collection.

*Recommendation: Update movie content regularly and consistently ensure that a plethora of newly released content is available. Assure that newly released movies are adequately advertised.*

One of the insights gained in the Analysis phase of the SOAR project insinuated that there was not only a higher frequency of rated content whose year of release was current to the data submission, but that these newly released films were often rated more favorably. Therefore, to increase user satisfaction and engagement, I suggested that an ample variety of consistently updated, current movie content be made available via the streaming service. Furthermore, it is imperative that current and potential users are aware of the new content’s availability by way of this figurative streaming service.

*Recommendation: Advertise the benefits of rating viewed movie content to users. Present an easy one-click rating popup at the conclusion of each movie viewing. Provide incentives for viewers to rate movie content.*

Another pattern gleaned from exploratory analysis was that much of the data was submitted by a relatively few number of users. This unfortunate observation could potentially introduce bias into the ratings data and obviously effect the accuracy. Consequently, all users must be encouraged to rate the movies after each viewing. Therefore, I suggested that users be made aware of the importance of rating movies to their user satisfaction. Furthermore, I suggested a ratings popup after every movie viewing with a convenient one-click rating questionnaire composed of simple, concise language to create a convenient, easy opportunity for all users to implement the desired behavior. Finally, some sort of incentive should be provided to users who engage with content by rating it. This incentive too would reinforce the desired behavior.

*Recommendation: Analysis highlights the SVD model as the most accurate. However, stakeholders should evaluate the advantages/disadvantages of each model in terms of computational expense, time to run each model, etc. Additionally, assess the tradeoffs between choosing an appropriate model according to the needs of the business immediately, or evaluating each model on different samples of data to contribute more evidence that the SVD model is in fact the most accurate.*

Stakeholders should assess the needs of the business and discern necessary aspects of a recommender system to address these needs. To elaborate, a free streaming service or one on the lower end of the cost distribution may find that the highest mean model or weighted mean model is sufficient for them, as a computationally expensive model may not be beneficial to them. This is especially true since there was little difference in accuracy across all the models. The business would need to decide if accuracy is the most desirable factor in a recommender system. Additionally, recall that the data set contains 100,000,000 observations. This amount of data requires processing power, computational expense, and professional talent with the ability to wield big data. Depending on the needs of the business, it may be beneficial to resample the data and run the models on different samples to test if the accuracy measures will be consistent with the findings above.

*Recommendation: Continue to update user-ratings data with current data collection.*

Current data and updated analysis ensure data trends will be spotted. The above discussion is a specified agenda implementing insights gleaned from the data, as well as recommendations for a plan of action. This discussion therefore exemplifies Objectives One, Three, and Five.

Once all the phases of the SOAR project were complete, it was necessary for me to effectively communicate the analysis, its findings, and resulting recommendations. If this step was unsuccessful, it could be the obstacle preventing necessary initiatives from implementation to the detriment of the company. This is likely the reason communication with different types of stakeholders within the business was seen to be important enough to be the requirement for Objective Six in the expected outcomes for the Applied Data Science program. Indeed, this import communication step informing relevant stakeholders within the business was one with which I received ample practice. Written communication about the project was realized in the form of a written report, complete with visualizations, and returns from models that we had built. The report was packed with relevant information required by related professional roles within the business. Additionally, the information was presented in language that could be understood by professionals outside the domain of data analytics. This can be evidenced by the written report included in the appendix section of this reflection and in various previous discussions on this SOAR project included in this very reflection. Through written communication, I included explanations on various statistical analyses and models. I did not assume that my audience were experts in statistics or analytics. Added to written communication of this nature, I also composed the presentation, and took part in verbally communicating this information which could have easily been understood by various related professional stakeholders in this figurative business. In fact, throughout the course of the IST 718 class, I shared the opportunity with my team members to present our progress check points multiple times over the course of the project. This provided the experience needed that would likely be required in a professional setting, as stakeholders would reasonably want to see the progress of any given project. The presentation slides associated with this SOAR project can also be found and viewed along with the written report in a file associated with this report. In this way the SOAR project enabled me to master Objective Six of the expected outcomes.

Finally, this project provided the opportunity for mastery of Objective Seven, which requires students to “synthesize the ethical dimensions of data science practice.” In alignment with this requirement, sensitive information describing customers was not provided in the data. Furthermore, customer identity was protected by the replacement of customer names with customer ID’s. Therefore, analysis was centered solely on customer movie preference and building the most accurate model. Preventing bias in analysis was slightly more challenging. This is due to the millions of observations included in the data. As mentioned before, neither my team members, nor myself owned a machine that was powerful enough to handle this magnitude of data efficiently. My assumption would be that in a professional setting, employees would be supplied with appropriate machines to handle the necessary data load required by the business. However, working on this project for a figurative professional setting, we did not have such access to appropriate machines. The result is that down sampling was required to perform analysis on this large quantity of data. Because this requirement presents an opportunity for bias, we randomly sampled the data. As discussed earlier, no additional information describing customers was provided in the data. Consequently, there was no way to cluster or classify users for stratified, or sampling strategies. Therefore, random sampling was the best attempt with the data at hand to prevent bias in the analyses. In these ways, my participation in this SOAR project displayed proficiency in Objective Seven.

With a professional background in the field of education, I would prefer to maintain a position in this domain as an educational analyst, then an educational data scientist. Although not entirely prevalent at this point in time, these positions do exist. Data scientists within the field of education are becoming more popular as their value is increasingly appreciated by educators who fancy themselves professionals driven by student data. With this ambition, I see and appreciate the potential value in building recommender systems and extrapolating this skill to student academic success. The models applied to user preference data surrounding movie content could easily be adapted and applied to recommending initiatives based on student response to previous implementations using assessment data. Additionally, with more information describing students, clustering could be utilized to form small groups using quantitative analysis. Currently small groups are often formed by teachers at the classroom level, using observational, as well as assessment data. However, quantitative analysis implemented in the formation of small instructional groups is rarely used according to my professional teaching experience. The use of clustering to form small groups, along with student response to instructional initiatives measured by assessment data could be used in recommender models to suggest effective instructional strategies/initiatives differentiated at the student cluster/small group level. These recommended strategies could then be implemented in small group instruction. Consequently, different small groups could receive simultaneous instruction targeted to their respective academic needs, optimizing one of educators’ most sparse resources…time.

INFLUENTIAL READINGS:

One of the primary readings that contributed to my newly acquired data science skillset was *Understanding Variation – The Key to Managing Chaos,* Wheeler (2000). This piece introduced me to control charts and the idea of discerning signals from noise. I specifically remember reading this book thinking that every super intendent from every school district in the country’s public school system, along with all school and district-based administrators should be required to read this book. I wondered how much more optimally the educational budget could be spent if public school educators understood that a given increase or decrease in student assessment data, relative to the very last assessment, could simply be natural variation. How many instructional initiatives were deemed ineffective because student assessment data decreased relative the previous academic year? These sort of high stakes decisions are made in many districts without any sort of analysis of variance or difference in means analysis, or even plotting the data on a control chart with upper and lower limits to see the variation over time. Refocusing on Wheeler (2000) not only was the book written in comprehensible language that requires no proficiency in statistics, but it directed the attention of the reader away from naive misconceptions about the practice of data-driven implementations. In this work, Wheeler specified the faults of focusing on the increase/decrease of data points at each temporal collection period, and instead revealed the importance of charting the data in a time-series fashion using upper and lower limits as a likely range in normalcy. The inference here is that data points venturing outside of these limits are possible contenders to be signals that something outside of the “normal” is happening and is worthy of further investigation. This book thoroughly enlightened me to this more appropriate and accurate consideration of data and therefore was a significant contribution to my fundamental understanding of data analysis.

The other work most notable work in building my foundational understanding of statistics was the book *Reasoning with Data – An Introduction to Traditional and Bayesian Statistics Using R*, Stanton (2017). This book thoroughly explains foundational concepts in statistics, from ground level explanations discussing measures of central tendency and dispersion, to progressively more sophisticated discussions explaining frequentist versus Bayesian modeling. Throughout the informational journey of reading this content, I was bombarded with the various applications that statistical analyses could have in education at the district, school, and even the classroom level. Applications such as time and resource optimization, as well as factors effecting student performance, and discerning the most effective instructional initiatives could be addressed using analyses described in this work. Up to this point in my academic journey, I had little familiarity with Bayesian thinking. Stanton (2017) made clear the idea of traditional frequentist testing inclusive of using p values communicating where the frequency of a particular happening would fall on a distribution, and the ability to utilize this estimation to decide whether to reject the null hypothesis describing a lack of statistical significance of some occurrence. He also clarified the concept of Bayesian statistics. This is a concept that seems to reveal much more information about some particular occurrence, in that a direct estimated range of probabilities describing some hypothesized happening within the actual population. This concept was a revelation for my understanding of creating models that would evaluate hypotheses.

In conclusion, projects and their associated courses described in this reflection, along with their facilitated opportunities for me to practice the objectives described as expected outcomes of this program, provided me with quantitative tools to accurately acquire answer to questions that a potential professional situation would likely pose. These projects, along with their associated coursework required skills that directly aligned with said objectives, as discussed throughout the reflection, and resultingly enabled me to gain proficiency in these standards. I will certainly be able to apply these newly acquired skills in my future professional setting.

REFERENCES

National Heart, Lung, and Blood Institute. *Calculate Your Body Mass Index.*[Calculate Your BMI - Standard BMI Calculator (nih.gov)](https://www.nhlbi.nih.gov/health/educational/lose_wt/BMI/bmicalc.htm)

Florida Department of Education. (2008-2019). *Florida Standards Assessments Results*.  
[Results (fldoe.org)](https://www.fldoe.org/accountability/assessments/k-12-student-assessment/results/)

United States Census Bureau. (2019). *Income and Poverty in the United States: 2019.*[Income and Poverty in the United States: 2019 (census.gov)](https://www.census.gov/library/publications/2020/demo/p60-270.html)

Crosnoe, Robert , and Cooper Carey E.. 2010. “Economically Disadvantaged Children's Transitions Into Elementary School: Linking Family Processes, School Contexts, and Educational Policy.” American Educational Research Journal 47 (2): 258–91. 10.3102/0002831209351564.

StackExchange. (2012). *Understanding the IMDb weighted rating function for usage on my own website.*  
[Understanding the IMDb weighted rating function for usage on my own website - Mathematics Stack Exchange](https://math.stackexchange.com/questions/169032/understanding-the-imdb-weighted-rating-function-for-usage-on-my-own-website)

<https://www.kaggle.com/laowingkin/netflix-movie-recommendation/data?select=qualifying.txt>

Vidiyala, R., (2020). How to Build a Movie Recommendation System. *Towards Data Science*. https://towardsdata science.com/ how-to-build-a-movie-recommendation-system-67e321339109

Wheeler, D. J. (2000). *Understanding variation – The key to managing chaos.* Spc Press.

Stanton, J. M. (2017). *Reasoning with data – An introduction to traditional and Bayesian statistics using R.* The Guilford Press.