FALL 2021

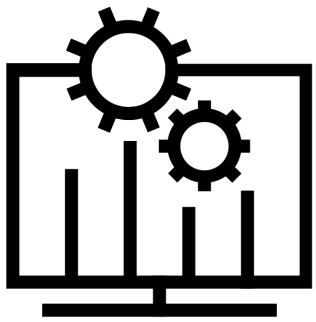
# MS APPLIED DATA SCIENCE PORTFOLIO MILESTONE

Rayanna Harduarsingh

Syracuse University

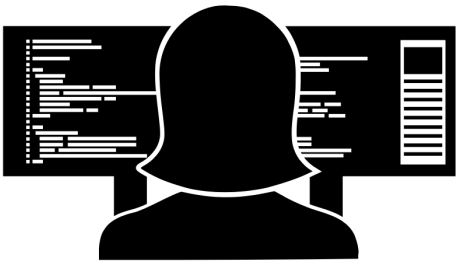
SUID: 224452222

Email: rharduar@syr.edu



# CONTENTS





# INTRODUCTION

As my time at Syracuse University draws to a close, I can't help but reflect on my academic career. I entered as an undecided, clueless freshman in 2016, and as a matter of fact, I ultimately graduated with two bachelor's degrees: Information Management & Technology and Advertising. While most of my peers were eager to venture out into the real world, I remained intrigued by higher education; I wanted to broaden my skills. Despite the fact that I studied two distinct subjects during my undergraduate studies, I've come to realize that they aren't totally different. In reality, technology is a tremendous benefit to advertising. The advertising business has evolved from a purely conventional model to one that is heavily reliant on consumer data. We are in a digital age and will be in one for a very long time. Technology is not only evolving and developing, but it is also revolutionizing industries, especially with the power of data. I wanted to pursue a Master's Degree in Applied Data Science to further explore the benefits and impacts of data analytics and how it can help businesses make better, more strategic business decisions and thrive as a company. This program has provided me with the technical and interpersonal skills essential to succeed in the field of data science. From collecting, analyzing, and visualizing data to translating it into actionable insights and putting them into practice through a plan of action, I can confidently apply for employment knowing that I am well-prepared. This milestone portfolio will demonstrate the following learning goals I have aimed to achieve of the Applied Data Science Master's program:

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

# INTRODUCTION TO DATA SCIENCE: IST 687

## COURSE DESCRIPTION

Led by Professor Jeffrey Saltz, this course provided me with a hands-on introduction to data science by exposing me to real-world examples of data collection, processing, transformation, management, and analysis. I was introduced to "R", a popular data analysis tool where I learned several data science concepts such as applied statistics, data visualization, text mining, and machine learning. At the end of the course, we were assigned a predictive analysis project to demonstrate our understanding of the class's learning objectives.

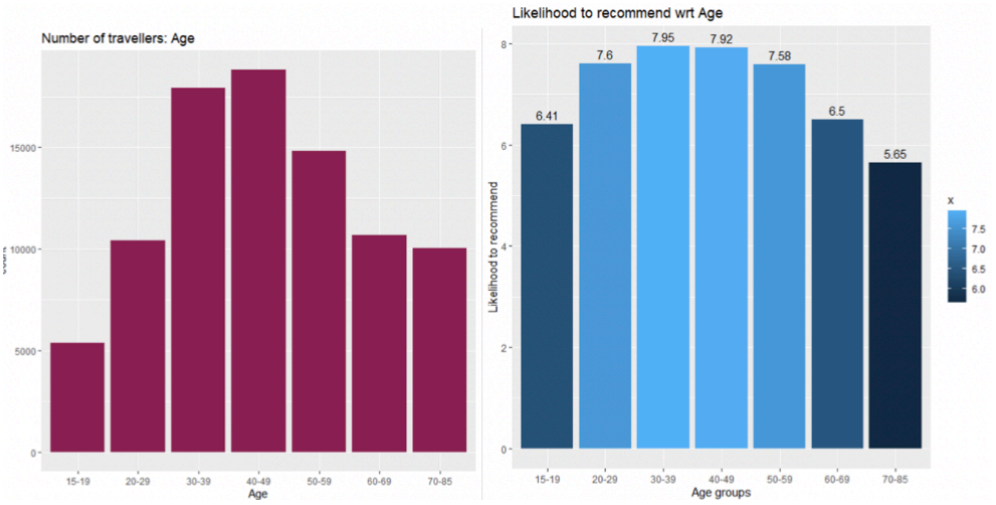
## ABSTRACT

A large survey was conducted by Southeast Airlines travelers who rated their overall satisfaction on a scale of 1-10, with 1 being the least satisfied, and 10 being the most satisfied.
We were given the task of improving Southeast Airline's customer churn by exploring, analyzing, and interpreting the findings of over 88,000 survey responses from travelers. We extensively evaluated significant factors as well as the travelers to predict which key factors have an influence on the satisfaction score. We discovered important trends in the data and provided Southeast Airlines with the best advice possible to increase customer satisfaction.

## DATA PREPARATION, METHODS, INSIGHTS, & RESULTS

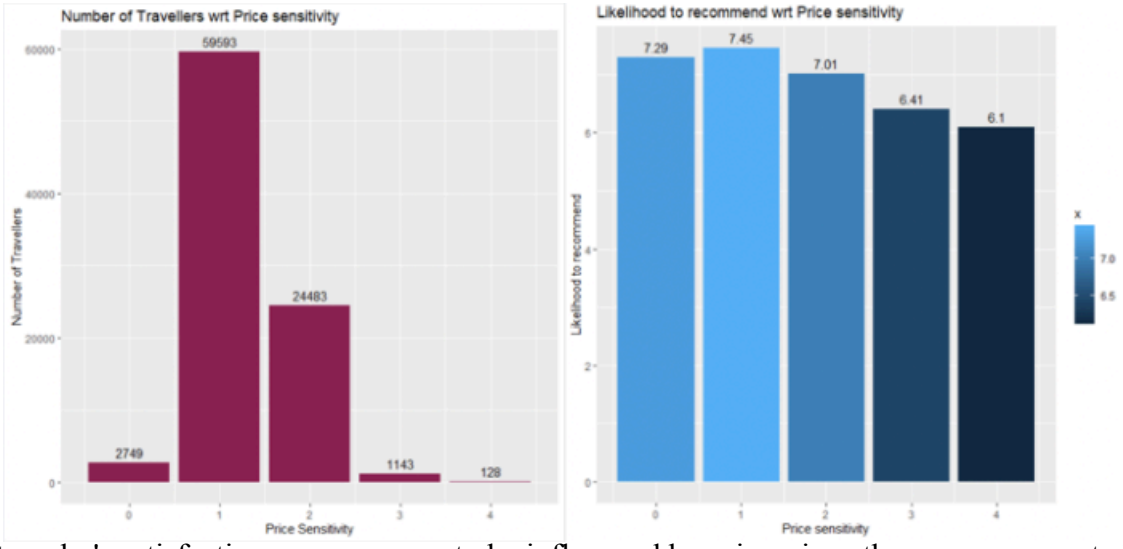
Using RStudio, we transformed the data format into a readable file where we then proceeded to clean the data, omit any missing values while keeping in mind a value of 0 did necessarily mean a blank field, and running descriptive statistics. We then translated the data into extensive visualizations to narrow down our analysis and choose which variables to explore further into. After exploring the data, we used four modeling techniques to create actionable insight: Association Rule Mining, Linear Modeling, Support Vector Modeling, and Text Mining. We were able to see which variables had a relationship with the survey rating and how it can influence the traveler's customer satisfaction.

## AGE



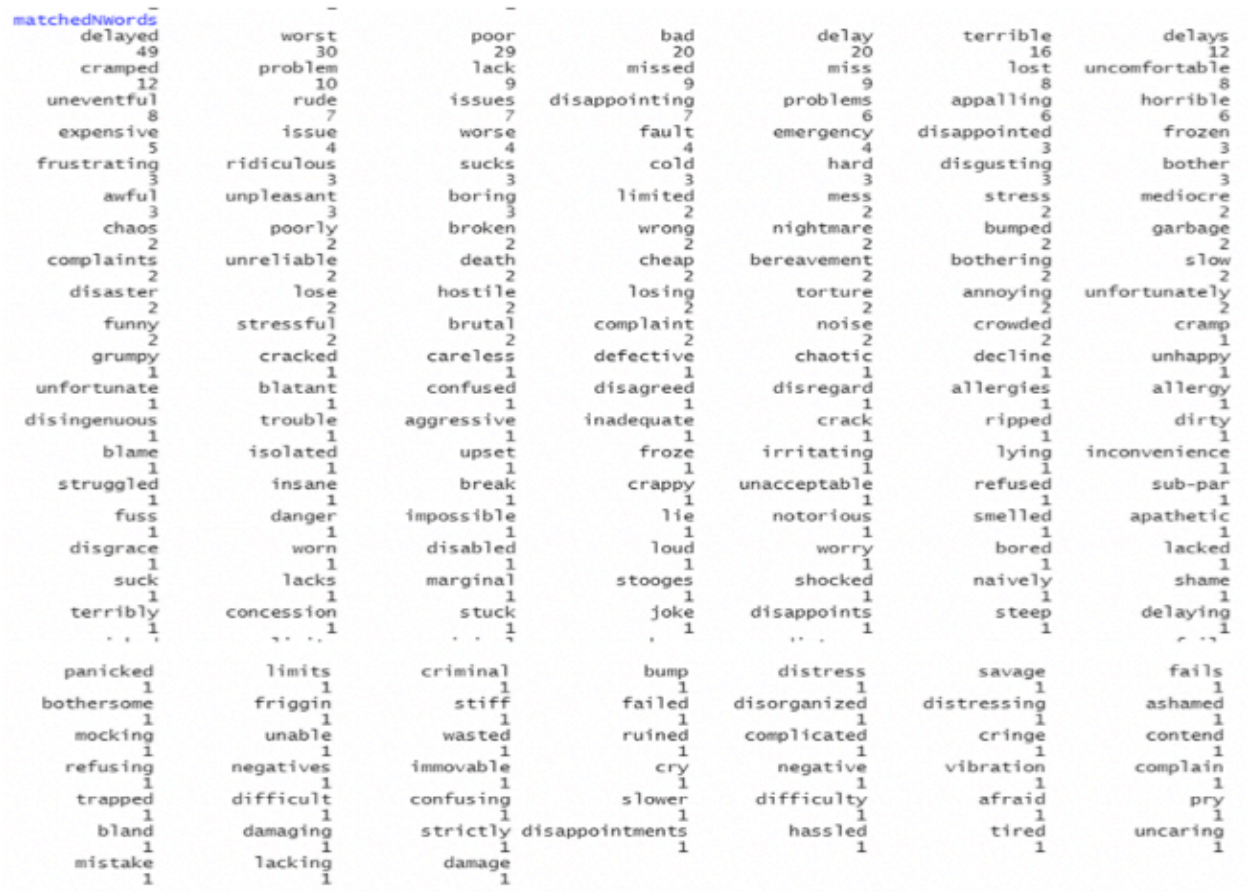

Age is always an important factor to begin with, so we looked to identify any patterns between certain age groups. Customers aged between 15-19 and 60-85 tended to give lower ratings, so this was a key variable looking further into. Using our other predictive models, we were given insight on how senior citizens should be extra-cared for such as offering blankets on their flights and how can better accommodate them to enjoy their trip.

## PRICE




The traveler's satisfaction score appears to be influenced by price, since the score appears to be greater when the price is low. This also gave us insight on possible recommendations for the airlines to offer seasonal promotions during special times of the year (Thanksgiving, Christmas, Summer, etc.).

## FLIGHT DELAYS



matchedNwords

| delayed | worst | poor | bad | delay | terrible | delays |
|---|---|---|---|---|---|---|
| 49 | 30 | 29 | 20 | 20 | 16 | 12 |
| cramped | problem | lack | missed | miss | lost | uncomfortable |
| 12 | 10 | 9 | 9 | 9 | 8 | 8 |
| uneventful | rude | issues | disappointing | problems | appalling | horrible |
| 8 | 7 | 7 | 7 | 6 | 6 | 6 |
| expensive | issue | worse | fault | emergency | disappointed | frozen |
| 5 | 4 | 4 | 4 | 4 | 3 | 3 |
| frustrating | ridiculous | sucks | cold | hard | disgusting | bother |
| 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| awful | unpleasant | boring | limited | mess | stress | mediocre |
| 3 | 3 | 3 | 2 | 2 | 2 | 2 |
| chaos | poorly | broken | wrong | nightmare | bumped | garbage |
| 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| complaints | unreliable | death | cheap | bereavement | bothering | slow |
| 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| disaster | lose | hostile | losing | torture | annoying | unfortunately |
| 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| funny | stressful | brutal | complaint | noise | crowded | cramp |
| 2 | 2 | 2 | 2 | 2 | 2 | 1 |
| grumpy | cracked | careless | defective | chaotic | decline | unhappy |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| unfortunate | blatant | confused | disagreed | disregard | allergies | allergy |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| disingenuous | trouble | aggressive | inadequate | crack | ripped | dirty |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| blame | isolated | upset | froze | irritating | lying | inconvenience |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| struggled | insane | break | crappy | unacceptable | refused | sub-par |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| fuss | danger | impossible | lie | notorious | smelled | apathetic |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| disgrace | worn | disabled | loud | worry | bored | lacked |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| suck | lacks | marginal | stooges | shocked | naively | shame |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| terribly | concession | stuck | joke | disappoints | steep | delaying |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| panicked | limits | criminal | bump | distress | savage | fails |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| bothersome | friggin | stiff | failed | disorganized | distressing | ashamed |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| mocking | unable | wasted | ruined | complicated | cringe | contend |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| refusing | negatives | immovable | cry | negative | vibration | complain |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| trapped | difficult | confusing | slower | difficulty | afraid | pry |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| bland | damaging | strictly | disappointments | hassled | tired | uncaring |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| mistake | lacking | damage | | | | |
| 1 | 1 | 1 | | | | |

Using Text Mining, we wanted to discover positive and negative word association in regards to the feedback. 178 negative words were matched and Delay was mentioned 80 times in total, which accounts for almost 50% of all negative words. From this analysis, we determined the word “delay” definitely factored into a negative rating. This gave us on insight for our business recommendation to accommodate travelers for flight delays such as complimentary snacks on their flight or a voucher for shopping at the airport.

## LEARNING GOALS

This was my first predictive analysis project, as well as my first time working with the programming language RStudio.When given a fresh data set, I learned how to run initial descriptive statistics to help us formulate business questions to aid in our analysis. It placed a strong focus on the ability to generate actionable insight from the data given, as well as the use of visualization and prediction models. While various machine learning and prediction

approaches were taught to us, it was critical that we filter down which techniques would be most effective. Prediction models were chosen depending on the data kind and the desired outcome, such as the ability to discover links and linkages that would provide relevant insights. Another significant learning objective was to be able to effectively communicate findings through our visualizations and analytics. Working with my group members who were in different parts of the world was a difficulty in and of itself since we were still in the thick of the pandemic. By adjusting to time differences and all-online sessions, communication was critical to the success of our project. Finally, in order to prevent prejudice and create unbiased insights, we had to apply ethics throughout our whole analysis. When working with data to make business decisions, ethics are just as crucial. To avoid public bias, we purposely avoided focusing on the gender variable. As a Data Scientist, it's critical to consider how public perceptions of your findings and suggestions will be reflected to avoid controversy.

# DATA ADMINISTRATION CONCEPTS & DATABASE MANAGEMENT: IST 659

## COURSE DESCRIPTION

Led by Professor Michael Fudge, this introductory course in database management systems taught me how to examine data structures, organize data, and implement data analysis using the Structured Query Language (SQL). By assessing business challenges and implementing effective data-oriented solutions, I learnt how to construct, model, and maintain databases as well as create queries. Advanced topics such as improving query efficiency for rapid analysis through indexing and evaluating options for data migrations, temporal data, and data standardization were also introduced to me. We were given a group project at the conclusion of the semester to demonstrate our ability to work as a team to design and implement a functional system with a database based on what we had learnt over the semester.

## ABSTRACT

The Hogwarts School of Witchcraft and Wizardry is entering the digital era and need assistance in digitizing and managing data about their school, students, and staff from a technological aspect in order to improve organizational efficiency. Our objective was to design a database management system for Hogwarts that would allow professors to efficiently manage and organize their institution's many departments in terms of students enrolled, classes, sports, and more. To provide a better user experience and faster information retrieval, all of this data was digitally stored.A user would be able to access student information such as their address for sending out school correspondence letters, what classes they are taking to ensure they are on schedule for graduation, the home they are allocated to, and even their pet that they are bringing for educational purposes. They can also use this information to plan their class schedule by seeing which professors are teaching certain classes. We've also entered wand information in case one breaks or disappears. Finally, we've added information on quidditch players so that faculty members may properly assign players to teams. Faculty members will be able to use our database system to ensure that all departments within their institution are running smoothly and to retrieve information more promptly.

## EXTERNAL DATA MODEL

### USER STORY: STUDENT CLASS ENROLLMENT

```
--As a faculty member, I should be able to see what classes a student is in so that they are on the right track to graduate.
select students.student_id,
    student_firstname + ' ' + student_lastname as student_name,
    student_year,
    classes.class_id,
    classes.class_name
from students
join student_classes on students.student_id=student_classes.student_id
join classes on student_classes.class_id=classes.class_id
```

Results Messages

| | student_id | student_name | student_year | class_id | class_name |
|---|---|---|---|---|---|
| 1 | 1 | Harry Potter | 5 | 1 | Defense Against the Dark … |
| 2 | 1 | Harry Potter | 5 | 3 | Astronomy |
| 3 | 1 | Harry Potter | 5 | 4 | Herbology |
| 4 | 1 | Harry Potter | 5 | 8 | Divination |
| 5 | 1 | Harry Potter | 5 | 9 | Care of Magical Creatures |
| 6 | 2 | Hermione Granger | 5 | 1 | Defense Against the Dark … |
| 7 | 2 | Hermione Granger | 5 | 3 | Astronomy |
| 8 | 2 | Hermione Granger | 5 | 4 | Herbology |
| 9 | 2 | Hermione Granger | 5 | 6 | Transfiguration |
| 10 | 2 | Hermione Granger | 5 | 8 | Divination |
| 11 | 2 | Hermione Granger | 5 | 9 | Care of Magical Creatures |
| 12 | 3 | Ronald Weasley | 5 | 1 | Defense Against the Dark … |
| 13 | 3 | Ronald Weasley | 5 | 3 | Astronomy |

## USER STORY 2: CLASS PROFESSORS

```
--As a student, I should be able to see what teacher is teaching a class so that I can make my schedule appropriate to my liking.
select class_id,class_name,faculty.faculty_id,
faculty_firstname + ' ' + faculty_lastname as Professor_name
from classes
join faculty on classes.faculty_id = faculty.faculty_id
```

| | class_id | class_name | faculty_id | Professor_name |
|---|---|---|---|---|
| 1 | 1 | Defense Against the Dark … | 2 | Severus Snape |
| 2 | 2 | Potions | 11 | Horace Slughorn |
| 3 | 3 | Astronomy | 9 | Gilderoy Lockhart |
| 4 | 4 | Herbology | 8 | Pomona Sprout |
| 5 | 5 | Dark Arts | 17 | Alastor Moody |
| 6 | 6 | Transfiguration | 3 | Minerva McGonagall |
| 7 | 7 | Music | 10 | Filius Flitwick |
| 8 | 8 | Divination | 15 | Sybill Trelawney |
| 9 | 9 | Care of Magical Creatures | 4 | Rubeus Hagrid |
| 10 | 10 | Charms and Spells | 5 | Dolores Umbridge |
| 11 | 11 | History of Magic | 6 | Remus Lupin |
| 12 | 12 | Flying | 12 | Rolanda Hooch |
| 13 | 13 | Muggle Studies | 14 | Cuthbert Binns |

## USER STORY 3: WAND INFORMATION

```
--As a student, I should be able to see the components of my wand so that if it ever goes missing I can know how to replace it.
select student_firstname + ' ' + student_lastname as student_name,
wands.wand_id,wand_length,wood_type,substance_name from wands
join wand_woods on wands.wand_id = wand_woods.wand_id
join wand_substances on wands.wand_id = wand_substances.wand_id
join woods on wand_woods.wood_id = woods.wood_id
join substances on wand_substances.substance_id = substances.substance_id
join students on wands.wand_id = students.wand_id
```

| | student_name | wand_id | wand_length | wood_type | substance_name |
|---|---|---|---|---|---|
| 1 | Harry Potter | 1 | 11.00 | Elm | Dragon Heartstring |
| 2 | Hermione Granger | 2 | 10.75 | Ash | Phoenix Feathers |
| 3 | Ronald Weasley | 3 | 12.00 | Hazel | Unicorn Hair |
| 4 | Ginerva Weasley | 4 | 9.50 | Red Oak | Veela Hair |
| 5 | Fred Weasley | 5 | 10.25 | English Oak | Thestral Tail Hair |
| 6 | George Weasley | 6 | 10.25 | Sugar Maple | Troll Whisker |
| 7 | Draco Malfoy | 7 | 10.00 | Pine | Kelpie Hair |
| 8 | Luna Lovegood | 8 | 11.25 | Snakewood | Thunderbird Tail Feather |
| 9 | Neville Longbottom | 9 | 12.25 | Chestnut | Wampus Cat Hair |
| 10 | Padma Patil | 10 | 9.00 | Reed | White River Monster Spine |

We created these sample user stories based off of the conceptual and logical models we developed to organize the relationships between students, faculty, classes, wands, wand length, wood type, and substances. These tables above were created to display the classes students are enrolled in, the professors teaching a class, and the components of each students' wand.

### LEARNING GOALS

What I found most interesting about this project was the liberty we were given to design our own database from scratch. For our external data model, my team and I had to actually create sample data. This demonstrated the capacity to collect, store, and access information through SQL. If there's one thing I've learned about data, it's how messy it can get. I learned how to properly clean and organize data so that analysis can be completed quickly and efficiently. Data management has taught me the importance of storing and accessing data in order to make business decisions and automate operational processes.

## DATA VISUALIZATION: IST 719

### COURSE DESCRIPTION

Led by Professor Jeff Hemsley, this course introduced me to a variety of skills and techniques for creating informational data visualizations. In RStudio's graphics environment, I learned how to clean data, create custom plots, and visually export data. In Adobe Illustrator, I also learned how to apply numerous design concepts to visually communicate a story told from the data. At the end of the course, we were assigned a final project to produce a poster that leveraged our skills developed throughout the semester. We were given the opportunity to work with a dataset of our choice to create several visualizations using data cleaning, visual exploration, aggregating variable, information organization, and design abilities.

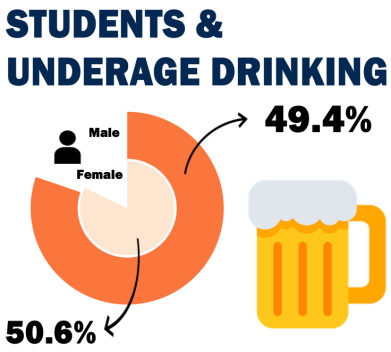
### ABSTRACT

Underage drinking has been a serious long-term issue in the United States and all over the world. According to the 2019 National Survey on Drug Use and Health, about 7.05 million Americans between the ages of 12 and 20 reported current alcohol consumption. A data set from 2008, provided by the University of California Irvine (UCI), obtained student data from a survey of a Math and Portuguese language course taught in a secondary school located in Portugal. It includes various social, gender, and study data from mark reports and questionnaires such as the student’s age, alcohol consumption, mother’s education, and more. The goal was to create

multiple visualizations that tells a story of how alcohol consumptions among students affect their academic activities.

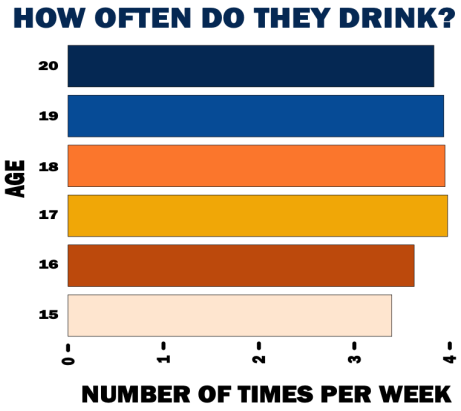
## EXPLORING THE DATA

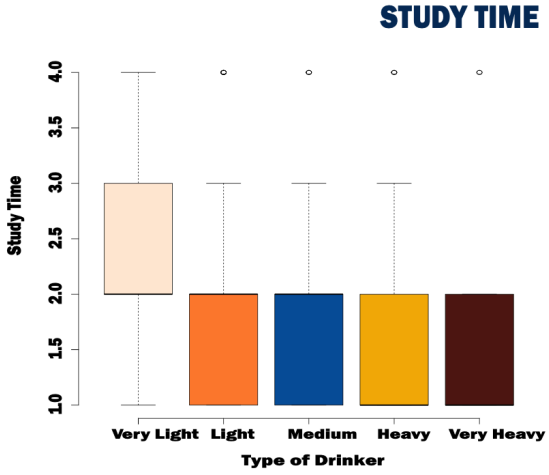
The overall story I wanted to tell was how does underage drinking affect a student's academic performance. First, I cleaned the data by combining the number of times a student drinks on the weekdays and weekend into one column to be used for the rest of my visual exploration. Then, I omitted rows of students who were over the age of 21 to report accurate results.



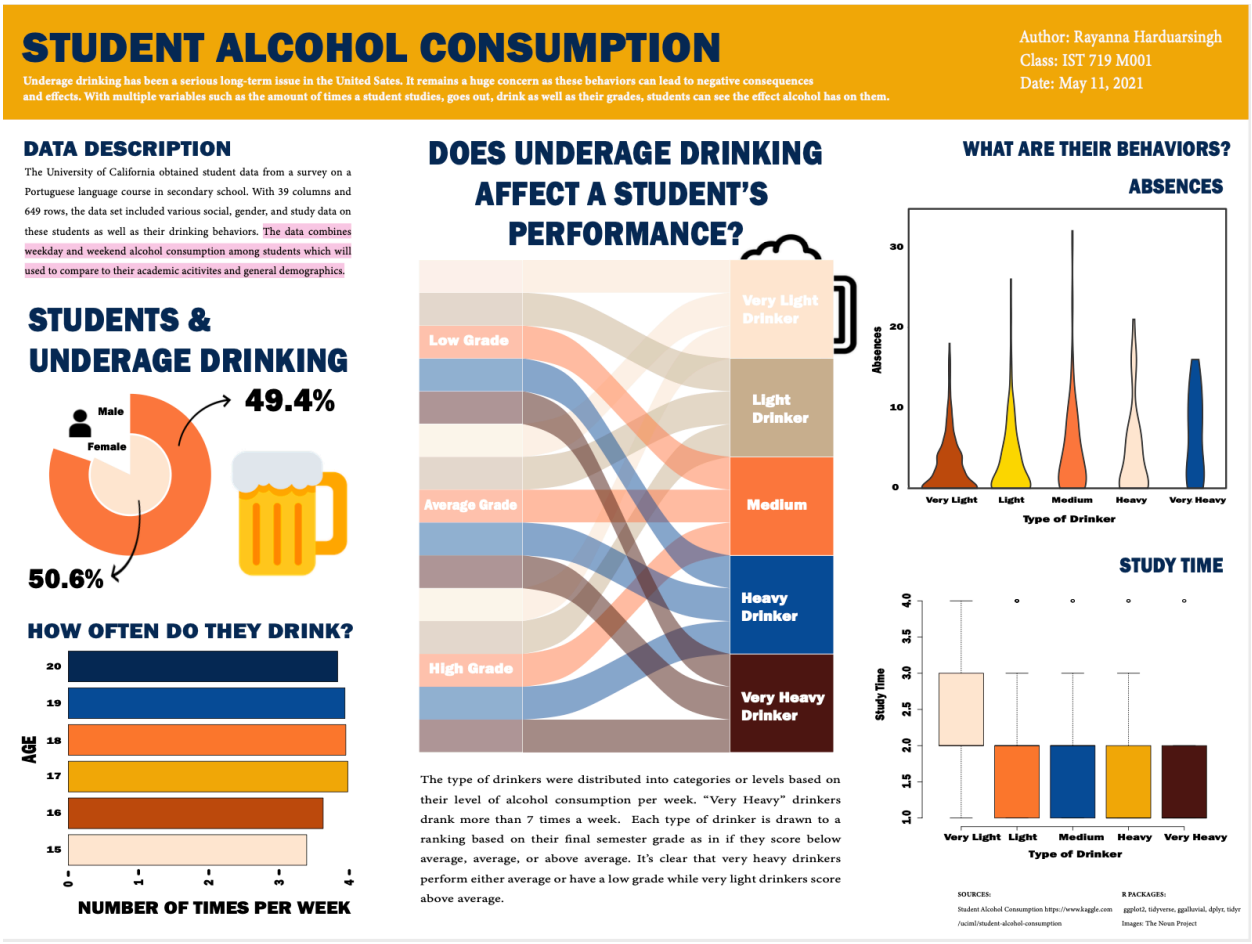

I first explored drinking by gender to see if being a male or female had a higher rate of drinking. I aggregated the drinking column by sex to display the sum of how many times each sex drinks and produced a layered donut visual. Females drank quite a bit more than males, however it was not such a drastic difference.

The next plot I explored was how often each age drinks per week. The lowest age in the dataset was 15 and the highest was 20 when omitting students who were over the age of 21. I, again, aggregated the data by the drinking column by age to display the average number of times each age drinks in a bar plot.




I also wanted to see what academic activities were affected by underage drinking. I explored the "study time" variable which includes the number of hours studied per week. I created a box plot and combined them into a single graph and we saw that the more the drinking, the less time a student studied.




I experimented with a violin and alluvial plot to easier classify the levels of drinkers. First, I organized the number of times a student drank per week into 5 different factor levels:

- **Very Light Drinker:** 1-2 times per week
- **Light Drinker:** 3-4 times per week
- **Medium Drinker:** 5-6 times per week
- **Heavy Drinker:** 7-8 times per week
- **Very Heavy Drinker:** 9-10 times per week

These graphs were able to tell the story of how underage drinking affects student's academic activities such as the amount of time they study, how many times they miss school, and most importantly, how it affects their final grades. Study time and absences definitely are affected as students study less and are absent the more they drink. However, we see that the distribution of drinking and final grades are a bit spread out, but we can slightly see that heavy drinkers either perform average or low in their grades.

### LEARNING GOALS

My final poster taught me how to communicate insights gained via visualization to a broad range of audiences. I learned there were much more descriptive plots we can create in RStudio than bar graphs, pie charts, and line graphs. One main takeaway that I found in this class how was important it is to tell a story from our data and how to communicate it to different audiences. Not everyone is a data scientist or analyst, so it's crucial to create a visual that any human eye can easily read. Once a visualization is made, the story should be able to be told in a few seconds. It's important to have clean, precise, and accurate data to produce an informational graphic. This allows for bias to be avoided and data not to be altered. It's not supposed to tell *your* story, but the data's story.

## DATA ANALYTICS: IST 707

### COURSE DESCRIPTION

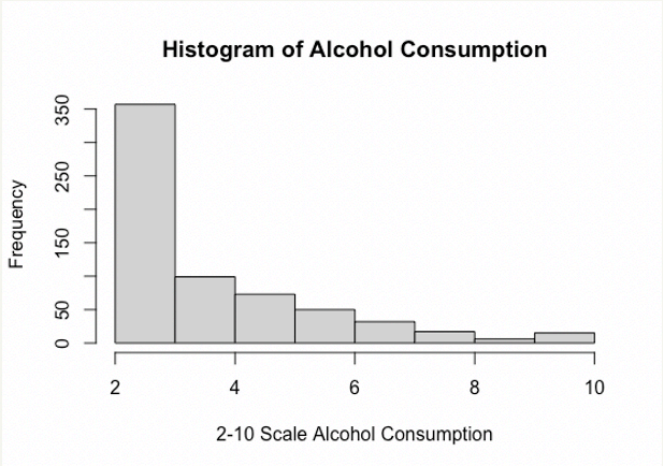
Led by Professor Yang Yang, this course introduced me to data analytics techniques, as well as experience with specific real-world applications, the issues associated with these applications, and the field's future directions. We got hands-on practice using open-source software packages in RStudio and Weka. We extracted insights from data using common data analytics approaches while studying the foundations and theories of data analytics methods and how to apply them to solve issues. We concentrated on how to comprehend data and create data analytics tasks in order to address problems with data. This course covered the fundamentals of data analytics, such as data preparation, concept definition, association rule analytics, classification, clustering, evaluation, and analysis. We learned abilities that may be applied to business, science, or other organizational challenges as a result of our research of data analytics principles and approaches, as well as practical activities. We completed a final data analysis report at the conclusion of the semester that highlighted a data analytics problem, its relevance and larger effect, data analytics methodologies, outcomes, and interpretation of the identified patterns.

## ABSTRACT

Student alcohol consumption is a prevalent public health issue, but little is known of its effects on academic achievement. Using real-world data obtained from a secondary school in Portugal, our objective was to predict how underage alcohol consumption affects a student's performance. To model our predictions, we used four different analysis methods: Logistic Regression to discover correlations, Clustering to segment students based on several metrics, Classification (Naive Bayes, Support Vector Machine, Decision Tree, & Random Tree) to predict accurate relationships, and Association Rule Mining to find common variables with students who had high or low alcohol consumption. Although the classification models did not show a good predictive accuracy, the regression and association rule mining yielded the most insights. Alcohol consumption did not heavily affect the grades of students, but it was more likely to be associated with higher absences, low study time, and failures. The results of this analysis could potentially identify what resources students need through student outreach and create policy to put an end to underage drinking.

## DATA DESCRIPTION & PRE-PROCESSING

The .csv data file we obtained from Kaggle contained 649 observations across 32 variables. For the purpose of this analysis, we combined two variables: Dalc (weekday alcohol consumption) and Walc (weekend alcohol consumption) into a single variable "alcohol". Dalc and Walc ranged 1-5, so the combined "alcohol" variable had values ranging from 2-10.




After combining them, the Dalc and Walc variables were removed as they would interfere with our classification algorithms. The right-skewed histogram below gives the number of students by their self-reported alcohol weekly alcohol consumption.

## METHODS & RESULTS

### LOGISTIC REGRESSION

We converted all of our character variables into factors and ran a logistic regression, with alcohol as our response variable and all other variables as the explanatory. We found that 11 of the 31 variables were significant at alpha = 0.05. The results of the regression can be seen below, surprisingly grades were not a significant predictor of alcohol consumption which was one of our initial predictions. The most notable predictors with the lowest p-values were: goout (1-5 range of how often a student went out), sexM (male student) and famrel (1-5 range of family relationship quality). Some results were expected, such as increased study time decreasing alcohol consumption, but there were some surprises such as health (1-5 range of health quality) increasing alcohol consumption which seems contradictory.

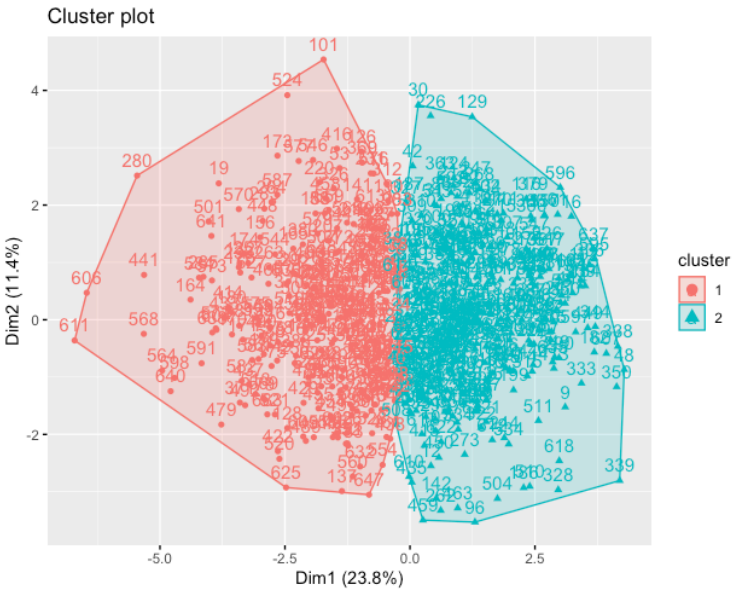
| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 1.076242 | 1.277970 | 0.842 | 0.4000 | |
| schoolMS | -0.063794 | 0.169926 | -0.375 | 0.7075 | |
| sexM | 1.082006 | 0.150160 | 7.206 | 1.72e-12 | *** |
| age | 0.107022 | 0.064006 | 1.672 | 0.0950 | . |
| addressU | -0.127898 | 0.162940 | -0.785 | 0.4328 | |
| famsizeLE3 | 0.327084 | 0.152108 | 2.150 | 0.0319 | * |
| PstatusT | 0.309054 | 0.215399 | 1.435 | 0.1519 | |
| Medu | -0.054828 | 0.093794 | -0.585 | 0.5591 | |
| Fedu | 0.088830 | 0.085305 | 1.041 | 0.2981 | |
| Mjobhealth | -0.301161 | 0.333710 | -0.902 | 0.3672 | |
| Mjobother | -0.203614 | 0.188394 | -1.081 | 0.2802 | |
| Mjobservices | -0.005160 | 0.232528 | -0.022 | 0.9823 | |
| Mjobteacher | 0.285325 | 0.312598 | 0.913 | 0.3617 | |
| Fjobhealth | 0.175200 | 0.467884 | 0.374 | 0.7082 | |
| Fjobother | 0.275100 | 0.284230 | 0.968 | 0.3335 | |
| Fjobservices | 0.631907 | 0.298267 | 2.119 | 0.0345 | * |
| Fjobteacher | -0.345741 | 0.419472 | -0.824 | 0.4101 | |
| reasonhome | 0.237116 | 0.176996 | 1.340 | 0.1809 | |
| reasonother | 0.525076 | 0.227985 | 2.303 | 0.0216 | * |
| reasonreputation | 0.146608 | 0.185461 | 0.791 | 0.4295 | |
| guardianmother | -0.260556 | 0.164728 | -1.582 | 0.1142 | |
| guardianother | -0.235689 | 0.329848 | -0.715 | 0.4752 | |
| traveltime | 0.087566 | 0.099148 | 0.883 | 0.3775 | |
| studytime | -0.194701 | 0.087292 | -2.230 | 0.0261 | * |
| failures | -0.058672 | 0.132068 | -0.444 | 0.6570 | |
| schoolsupyes | 0.006324 | 0.229614 | 0.028 | 0.9780 | |
| famsupyes | 0.009443 | 0.141913 | 0.067 | 0.9470 | |
| paidyes | 0.228361 | 0.287537 | 0.794 | 0.4274 | |
| activitiesyes | -0.038275 | 0.139074 | -0.275 | 0.7832 | |
| nurseryyes | -0.356707 | 0.168281 | -2.120 | 0.0344 | * |
| higheryes | -0.006159 | 0.242508 | -0.025 | 0.9797 | |
| internetyes | 0.073936 | 0.171955 | 0.430 | 0.6674 | |
| romanticyes | 0.137881 | 0.142824 | 0.965 | 0.3347 | |
| famrel | -0.326104 | 0.071299 | -4.574 | 5.81e-06 | *** |
| freetime | -0.042925 | 0.069817 | -0.615 | 0.5389 | |
| goout | 0.588733 | 0.061536 | 9.567 | < 2e-16 | *** |
| health | 0.121688 | 0.047923 | 2.539 | 0.0114 | * |
| absences | 0.048483 | 0.015462 | 3.136 | 0.0018 | ** |
| G1 | -0.033363 | 0.050370 | -0.662 | 0.5080 | |
| G2 | 0.031612 | 0.065922 | 0.480 | 0.6317 | |
| G3 | -0.055573 | 0.053814 | -1.033 | 0.3022 | |

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

### CLUSTERING

We subsetted our initial dataframe into a new object “numdata” which only included numerical variables in order to be compatible with k-means clustering. After scaling the data, we determined our optimal number of clusters to be 2, as can be seen from the “elbow” in the wss plot below. After generating our 2 clusters, we can see from the cluster plot that there is some overlap between the two. When the number of clusters were increased to 3 and 4 we saw significantly more overlap and decided to continue with 2 clusters. Using the colMeans()




command on our "numdata" object with our clusters nested in brackets, we were able to view the average values of the numeric variables in the two clusters. In the first cluster (size 277), the average age, alcohol consumption, absences and failures were above the data set average, while the second (size 372) had lower than average values for the same variables. The grades for the first three quarters were lower on average for the first cluster, and higher for the second.

## CLASSSIFICATION

To test the predictive performance of alcohol consumption, we used the algorithms Support Vector Machine, Naïve Bayes, Decision Tree, and Random Forest to build our models. We split 70% of the data to use to train our prediction models, and the remaining 30% was left for us to test our models on. We also ran a 3-fold cross-validation procedure across each prediction model. Initially, we trained our models with alcohol as our factor variable in response to all the other explanatory variables. We were getting pretty low accuracies, even after predicting on our test set. In addition, we paired alcohol with other explanatory variables one at a time such as absences, final grades, failures, and more to predict a relationship, but we were still deriving low accuracies in the 9%-20% range. However, in order to improve the accuracy, we tried a different approach which would ultimately create our final prediction models. Based on our regression model from earlier, we identified quite a few statistically significant predictors that had a low p-value. We chose to use the top 4 predictors as our explanatory variables in our training models which were free time, going out, absences, and study-time in response to alcohol. For our SVM model, we tuned the parameter for our kernel to use the "polynomial" method instead of the classic linear as it produced a higher accuracy. For naïve bayes, we trained it using the default

naïve bayes method. For decision tree, we used the "rpart" method and for random forest we used the "rf" method. Each model was tuned with a search grid as well multiple different metrics to try and derive the highest accuracy as possible. It was then used to predict the testing set in the output of a confusion matrix. After using the statistically significant predictors from our regression, our accuracies definitely increased, however, it was still not a result we were looking for, hence trying other methods.

| MODEL | PREDICTION ACCURACY |
|---|---|
| SVM | 39.5% |
| NAÏVE BAYES | 37.9% |
| DECISION TREE | 34% |
| RANDOM FOREST | 32% |

## ASSOCIATION RULE MINING

We added an identity variable "ID" with the respective observation numbers to our data-frame, then used the mutate_if() function to convert all character variables into factors and discretize all numeric variables. By changing the data types, we were able to use the apriori() function on the data-frame in order to generate association rules. With roughly 3.78 being the average alcohol consumption across the entire dataset, we set the left-hand side apriori parameter to a combined list of alcohol values 5-10 for above average alcohol consumption. With our initial parameters set to 90% confidence, support equal to 0.001 and maximum length of 5, close to 3 thousand rules were generated. After inspecting the counts, we noticed that a lot of the rules had counts equal to 1. By filtering the rules by count numbers more than 3, we were left with 13 good rules which can be seen in the screenshot below. This process was repeated for the left hand side set to alcohol = 2 for below average/no alcohol consumption (26 good rules), along with higher (23 rules) and lower absence (18 rules) numbers. For absences, higher count and support filters had to be used in order to obtain the final results.

| | lhs | | rhs | support | confidence | coverage | lift | count |
|---|---|---|---|---|---|---|---|---|
| [1] | {reason=home,guardian=other,freetime=[4,5],absences=[4,32]} | => | {alcohol=6} | 0.006163328 | 1 | 0.006163328 | 12.980000 | 4 |
| [2] | {reason=home,guardian=other,romantic=no,goout=[4,5]} | => | {alcohol=6} | 0.006163328 | 1 | 0.006163328 | 12.980000 | 4 |
| [3] | {sex=M,reason=home,guardian=other,absences=[4,32]} | => | {alcohol=6} | 0.006163328 | 1 | 0.006163328 | 12.980000 | 4 |
| [4] | {reason=home,guardian=other,famrel=[4,5],absences=[4,32]} | => | {alcohol=6} | 0.006163328 | 1 | 0.006163328 | 12.980000 | 4 |
| [5] | {famsize=GT3,reason=home,guardian=other,romantic=no} | => | {alcohol=6} | 0.006163328 | 1 | 0.006163328 | 12.980000 | 4 |
| [6] | {Pstatus=T,reason=home,guardian=other,romantic=no} | => | {alcohol=6} | 0.006163328 | 1 | 0.006163328 | 12.980000 | 4 |
| [7] | {Mjob=health,famrel=[1,4),goout=[4,5],G2=[13,19]} | => | {alcohol=5} | 0.006163328 | 1 | 0.006163328 | 8.890411 | 4 |
| [8] | {Mjob=health,romantic=no,famrel=[1,4),G3=[13,19]} | => | {alcohol=5} | 0.006163328 | 1 | 0.006163328 | 8.890411 | 4 |
| [9] | {Fedu=[3,4],Mjob=health,traveltime=[2,4],goout=[4,5]} | => | {alcohol=5} | 0.006163328 | 1 | 0.006163328 | 8.890411 | 4 |
| [10] | {address=U,Fedu=[3,4],Mjob=health,traveltime=[2,4]} | => | {alcohol=5} | 0.006163328 | 1 | 0.006163328 | 8.890411 | 4 |
| [11] | {Medu=[0,2),reason=home,absences=[4,32],G2=[0,10)} | => | {alcohol=6} | 0.006163328 | 1 | 0.006163328 | 12.980000 | 4 |
| [12] | {Fedu=[3,4],famrel=[1,4),goout=[4,5],G3=[0,11)} | => | {alcohol=6} | 0.007704160 | 1 | 0.007704160 | 12.980000 | 5 |
| [13] | {Mjob=services,guardian=father,famsup=no,activities=no} | => | {alcohol=5} | 0.007704160 | 1 | 0.007704160 | 8.890411 | 5 |

| | lhs | | rhs | support | confidence | coverage | lift | count |
|---|---|---|---|---|---|---|---|---|
| [1] | {schoolsup=yes,nursery=yes,goout=[1,3)} | => | {alcohol=2} | 0.02927581 | 0.9047619 | 0.03235747 | 2.436475 | 19 |
| [2] | {traveltime=[1,2),schoolsup=yes,nursery=yes,goout=[1,3)} | => | {alcohol=2} | 0.02003082 | 0.9285714 | 0.02157165 | 2.500593 | 13 |
| [3] | {school=GP,sex=F,schoolsup=yes,goout=[1,3)} | => | {alcohol=2} | 0.02157165 | 0.9333333 | 0.02311248 | 2.513416 | 14 |
| [4] | {sex=F,famsize=GT3,schoolsup=yes,goout=[1,3)} | => | {alcohol=2} | 0.02003082 | 0.9285714 | 0.02157165 | 2.500593 | 13 |
| [5] | {sex=F,schoolsup=yes,nursery=yes,goout=[1,3)} | => | {alcohol=2} | 0.02157165 | 0.9333333 | 0.02311248 | 2.513416 | 14 |
| [6] | {schoolsup=yes,famsup=yes,romantic=no,goout=[1,3)} | => | {alcohol=2} | 0.02003082 | 0.9285714 | 0.02157165 | 2.500593 | 13 |
| [7] | {schoolsup=yes,famsup=yes,nursery=yes,goout=[1,3)} | => | {alcohol=2} | 0.02311248 | 0.9375000 | 0.02465331 | 2.524637 | 15 |
| [8] | {schoolsup=yes,nursery=yes,romantic=no,goout=[1,3)} | => | {alcohol=2} | 0.02311248 | 0.9375000 | 0.02465331 | 2.524637 | 15 |
| [9] | {famsize=GT3,schoolsup=yes,nursery=yes,goout=[1,3)} | => | {alcohol=2} | 0.02465331 | 1.0000000 | 0.02465331 | 2.692946 | 16 |
| [10] | {schoolsup=yes,nursery=yes,higher=yes,goout=[1,3)} | => | {alcohol=2} | 0.02773498 | 0.9000000 | 0.03081664 | 2.423651 | 18 |

| | lhs | | rhs | support | confidence | coverage | lift | count |
|---|---|---|---|---|---|---|---|---|
| [1] | {Medu=[0,2),paid=no,G1=[13,19]} | => | {absences=[0,4)} | 0.04314330 | 0.9032258 | 0.04776579 | 1.571564 | 28 |
| [2] | {Pstatus=T,Medu=[0,2),paid=no,G1=[13,19]} | => | {absences=[0,4)} | 0.04160247 | 0.9000000 | 0.04622496 | 1.565952 | 27 |
| [3] | {Medu=[0,2),paid=no,higher=yes,G1=[13,19]} | => | {absences=[0,4)} | 0.04314330 | 0.9032258 | 0.04776579 | 1.571564 | 28 |
| [4] | {Medu=[0,2),failures=[0,3],paid=no,G1=[13,19]} | => | {absences=[0,4)} | 0.04314330 | 0.9032258 | 0.04776579 | 1.571564 | 28 |
| [5] | {Medu=[0,2),paid=no,romantic=no,alcohol=2} | => | {absences=[0,4)} | 0.04160247 | 0.9000000 | 0.04622496 | 1.565952 | 27 |
| [6] | {school=MS,guardian=father,activities=no,famrel=[4,5]} | => | {absences=[0,4)} | 0.04314330 | 0.9032258 | 0.04776579 | 1.571564 | 28 |
| [7] | {school=MS,Pstatus=T,guardian=father,activities=no} | => | {absences=[0,4)} | 0.04930663 | 0.9142857 | 0.05392912 | 1.590808 | 32 |
| [8] | {school=MS,guardian=father,activities=no,higher=yes} | => | {absences=[0,4)} | 0.04468413 | 0.9354839 | 0.04776579 | 1.627692 | 29 |
| [9] | {school=MS,guardian=father,schoolsup=no,activities=no} | => | {absences=[0,4)} | 0.04622496 | 0.9090909 | 0.05084746 | 1.581769 | 30 |
| [10] | {school=MS,sex=F,guardian=father,famrel=[4,5]} | => | {absences=[0,4)} | 0.04468413 | 0.9354839 | 0.04776579 | 1.627692 | 29 |

| | lhs | | rhs | support | confidence | coverage | lift | count |
|---|---|---|---|---|---|---|---|---|
| [1] | {guardian=other,internet=yes,G2=[10,13)} | => | {absences=[4,32]} | 0.02003082 | 0.9285714 | 0.02157165 | 2.183489 | 13 |
| [2] | {famsize=GT3,Mjob=other,guardian=other,schoolsup=no} | => | {absences=[4,32]} | 0.02003082 | 0.9285714 | 0.02157165 | 2.183489 | 13 |
| [3] | {guardian=other,failures=[0,3],internet=yes,G2=[10,13)} | => | {absences=[4,32]} | 0.02003082 | 0.9285714 | 0.02157165 | 2.183489 | 13 |
| [4] | {school=GP,Mjob=teacher,studytime=[1,2),health=[3,5]} | => | {absences=[4,32]} | 0.02003082 | 0.9285714 | 0.02157165 | 2.183489 | 13 |
| [5] | {Mjob=services,guardian=mother,paid=no,nursery=no} | => | {absences=[4,32]} | 0.02003082 | 0.9285714 | 0.02157165 | 2.183489 | 13 |
| [6] | {age=[17,22],address=U,nursery=no,G3=[0,11)} | => | {absences=[4,32]} | 0.02311248 | 0.9375000 | 0.02465331 | 2.204484 | 15 |
| [7] | {sex=M,traveltime=[2,4],nursery=no,G2=[10,13)} | => | {absences=[4,32]} | 0.02311248 | 0.9375000 | 0.02465331 | 2.204484 | 15 |
| [8] | {Pstatus=T,Mjob=services,Fjob=other,G3=[11,13)} | => | {absences=[4,32]} | 0.02003082 | 0.9285714 | 0.02157165 | 2.183489 | 13 |
| [9] | {school=GP,schoolsup=no,goout=[1,3),G2=[0,10)} | => | {absences=[4,32]} | 0.02003082 | 0.9285714 | 0.02157165 | 2.183489 | 13 |
| [10] | {reason=home,internet=yes,famrel=[1,4),G1=[10,13)} | => | {absences=[4,32]} | 0.02157165 | 0.9333333 | 0.02311248 | 2.194686 | 14 |

## RESULTS & INSIGHTS

Out of all the methods used for our analysis, the regression model and association rules yielded the most insightful analysis. We were able to draw appropriate conclusions about the effects of student alcohol consumption. Underage drinking does affects students, but not dramatically based off of this data set. Our hypothesis that higher drinking levels will significantly correlate

with low final grades could not be proven as there was not much of a correlation or relationship between alcohol and grades. However, we were able to answer our second research question, “Can other variables besides alcohol consumption affect student performance?”. Alcohol consumption was found to be more associated with higher absences, low study time, and failures, which are all key components to have a good academic standing. Although some of our models did not achieve high accuracies on testing sets, we were still able to identify the significant variables that contributed to higher rates of alcohol consumption. After performing our analysis and coming to our conclusions, we reflected on how our analysis could have been more promising. We noticed some data observations were repetitive having several of the same values which is why we believe our accuracies were low in our prediction models from our classification method. However, being that it derived low accuracies, it could have that alcohol consumption was simply not a strong predictor for student performance. Having more observations in the dataset could potentially increase the accuracies of our models while also yielding more interesting association rules.

## LEARNING GOALS

This data analysis project taught me how to create actionable insight across a range of contexts using data. I was able to define, evaluate, and transform the issues related to data analytics into technological designs and solutions. I applied real-world challenges to data analytics principles, algorithms, and assessment methodologies. To employ data storytelling, I was able to experiment with several methodologies such as regression, classification, and machine learning methods, therefore demonstrating the ability to apply predictive models to help generate actionable insight. These strategies are critical for delving into data, discovering relevant patterns, and communicating what patterns have been discovered, how they were discovered, and why they are important and trustworthy.

## CONCLUSION

Overall, I've gained a great deal of useful skills that will undoubtedly be useful in the real world. These are the kinds of abilities that will help me advance in my career. I'm more prepared and confident than ever to begin my profession. From our homework to projects, we dealt with hands-on data and analytic techniques that made you feel like a real Data Scientist. I strongly believe I have fulfilled all six learning goals of the Applied Data Science Master's program.

### 1. COLLECT, STORE, AND, ACCESS DATA BY IDENTIFYING AND LEVERAGING APPLICABLE TECHNOLOGIES

Being able to collect, store, and access data by identifying and leveraging applicable technologies was demonstrated in course IST 659. In IST 659, we were given the task of creating data from scratch to collect and store it into a database management system using SQL. We then had to create multiple user stories or scenarios in which a user would access the system and how they would benefit from this data being transformed into a database.

### 2. CREATE ACTIONABLE INSIGHT ACROSS A RANGE OF CONTEXTS USING DATA AND THE FULL DATA SCIENCE LIFE CYCLE

Being able to create actionable insight across a range of contexts using data and the full data science life cycle was demonstrated in course IST 687 and IST 707. We went through the entire data science life cycle as part of the survey data analysis project and the student alcohol consumption analysis. From a business standpoint, we recognized a problem such as how can we reduce customer churn or how does alcohol affect a student's academic performance? The respective data was then collected, cleaned, and prepared. After cleaning the data, we can use statistics and visualizations to undertake exploratory data analysis, which will lead to our prediction models. We can then test our prediction models and create actionable insights, which will lead to the deployment of business recommendations.

### 3. APPLY VISUALIZATION AND PREDICTIVE MODELS TO HELP GENERATE ACTIONABLE INSIGHT

Being able to apply visualization and predictive models to help generate actionable insight was demonstrated in ST 687, IST 719, and IST 707. Visualizations are critical for breaking down large data sets, as taught in IST 719; they're a critical tool for making data more understandable to the human eye. We can find trends and patterns in visuals that can help us construct prediction models. Predictive analysis is performed with machine learning models that helps generate possible predictions for the future results. It involves using historical data and once validated, it is subsequently used to forecast what will happen next using current data. We can predict what groceries a certain person is likely to buy next or identify spam versus non-spam in e-mail systems. I also studied how linear regressions and machine learning models may assist in the identification of relationships between two or more variables in IST 687 and IST 707. We may look for correlations and find which major factors influence each other the most. We witnessed, for example, how a traveler's bad rating is influenced by delayed flights. In addition, even though we could prove our hypothesis in our student alcohol consumption dataset, we were still able to discover what factors into poor academic performance from our prediction models. When it comes to identifying links and generating actionable knowledge, predictive models are incredibly useful.

## 4. USE PROGRAMMING LANGUAGES SUCH AS R AND PYTHON TO SUPPORT THE GENERATION OF ACTIONABLE INSIGHT

RStudio is a powerful statistical analysis environment with a wide range of capabilities. Predictive modeling, machine learning methodologies, text mining, and creating visual plots from a data collection were among the skills I learned. These were all useful ways for extracting insights and informations from a data set using its results. In course IST 718, I'm currently studying how to utilize Python to get meaningful insights using analytic techniques from a variety of Python packages.

## 5. COMMUNICATE INSIGHTS GAINED VIA VISUALIZATION AND ANALYTICS TO A BROAD RANGE OF AUDIENCES

This learning goal exposed me to the different focus areas of data science. I enjoyed making visualizations from a large data set the most since it taught you how to make a visual from multiple viewpoints. Not everyone can be a data scientist, and statistics might be difficult to

comprehend. This is why visualizations are important when presenting your findings in order to create an impression on a wide range of people and make them understandable. When developing a visualization, one of the most essential things to remember is to convey a story. The story should flow smoothly and be clearly visible. The first step in developing a visualization is to do adequate analytics. Then, when it comes to analytics, it's critical to choose the correct style of graph to represent those statistics. Then comes the difficult part: designing it. Colors and colors play an equal role in your discoveries. When using visualizations to communicate your ideas, make sure your image isn't cluttered. It must be appropriately aligned and arranged, with the suitable colors and graph playing a significant role.

## 6. APPLY ETHICS IN THE DEVELOPMENT, USE AND EVALUATION OF DATA AND PREDICTIVE MODELS

Finally, my ability to apply ethics in the development, usage, and evaluation of data and prediction models was exhibited throughout this program in all of my classes. When it comes to analyzing data and developing prediction models, ethics is crucial. In order to avoid disseminating false information and misinforming the public, you must consider how accurate your analysis and insights are. For example, in its 719, I needed to be sure I was only dealing with underage drinkers, thus it was critical that I omitted any rows that had students over the age of 21 from my analysis in order to get an accurate result. When working with sensitive matters like gender, you want to avoid any bias in your study as a data scientist. The public may interpret the outcomes as biased. You also don't want to change the data in any manner to get a conclusion that is influenced by your or others' opinions. It's critical to practice fairness in your analysis and consider how the public could react to your findings and conclusions.

As I near the completion of my Applied Data Science Master's degree, I am still learning so much more, demonstrating all six of these learning objectives. In my present classes, I continue to use what I've learnt in the past. I can see how data scientists play a critical role in our digital age, and how we can influence the future of businesses in a variety of industries. I look forward to my journey post-graduation and I am forever grateful for the Applied Data Science program.

**Go ‘Cuse! - A Forever Orange**