Syracuse University School of Information Studies
M.S. Applied Data Science

M.S. in Applied Data Science
Final Program Portfolio

Syracuse University - Summer 2021

Samuel Bull
SUID#: 984005814
stbull@syr.edu

GitHub Link: https://github.com/stbull/Syracuse_MSADS_Portfolio

Video Presentation Link: https://youtu.be/wF4FjpNIBYE

Table of Contents

Introduction:

At Syracuse University, the Master's program in Applied Data Science through the School of Information Studies allows its students to learn different techniques of collecting, analyzing, and finding solutions while working with many types of data coming from various sources. In today's day and age, it is becoming more and more apparent that having the skills to work with data is an integral part of the modern workplace. This program and its courses, such as an Introduction to Data Science, Data Administration Concepts and Database Management, Big Data Analytics, and Information Visualization have allowed me to both continue building upon skills I began learning during my undergraduate collegiate career, as well as create new ones. With the skills I've learned through programs and coding languages like Microsoft Excel and Access, R-Studio, and Python (to name a few), I'm confident that I now have the ability to be an insightful, knowledgeable, and detail-oriented member of any career path I may take moving forward.

The iSchool at Syracuse University has set seven different learning goals that should be achieved by people in the ADS program. Those seven goals are:

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)

Throughout this portfolio, I attempt to show how I have applied my skills to be able to fulfill these objectives in the different courses and projects I've completed during my time within the program.

Applications:

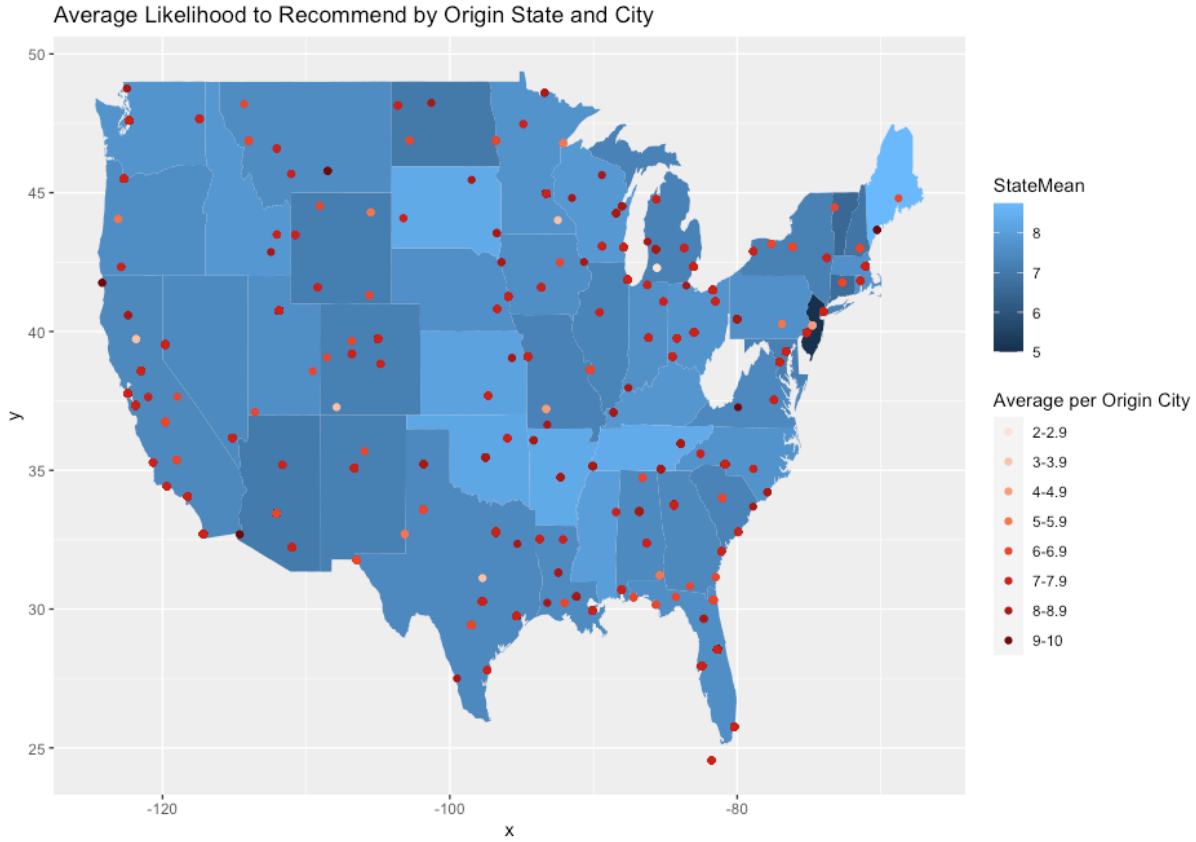
IST 687: Introduction to Data Science

Project Overview:

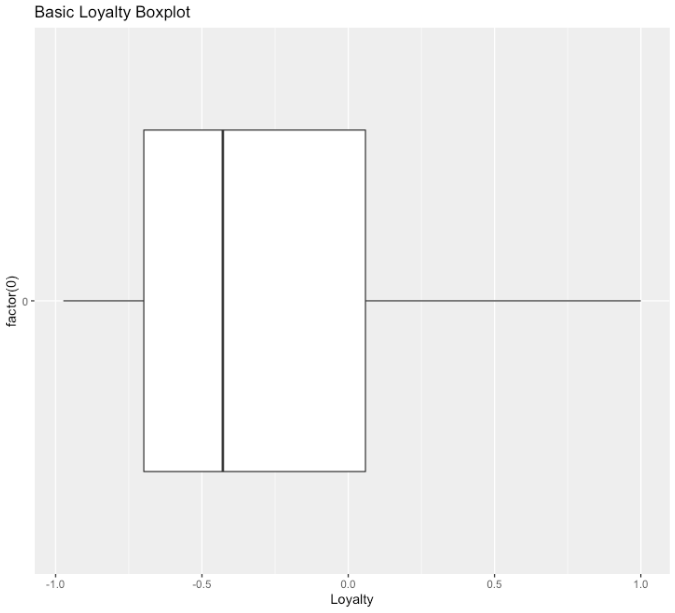
Throughout this Introduction to Data Science course, which I took asynchronously under the guidance of Professor Jeffery Saltz, I began to generate skills in R and R-Studio; a software that I would use across several different courses I took as a part of this program. From the lessons that I learned, I was tasked with taking data that dealt with fliers on a fictional airline to analyze and pull insights from with the goal of determining if a passenger would be likely to recommend the airline or not. While working with the data, I was able to hone my skills when it came to everything from data visualization (utilizing packages ggplot2 and maps), association rules data mining, frequentist (linear model) regression, and classification through different Support Vector Machines.

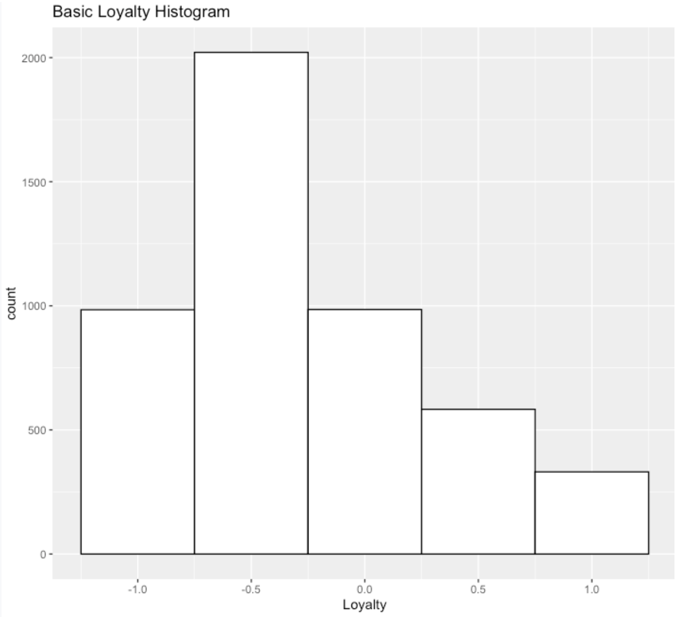
Because of the analysis I performed, I was then able to create solutions that I saw fit for fixing any problems that came forth through the basic regression models I created. For my final presentation of the semester, a PowerPoint presentation was created to both showcase my findings, explain the way I went about creating my regression and classification models the way I did, and give any suggestions to fix the issues that I set out to resolve for the fictional airline.

Screenshot Examples from Project:



Map of Likelihood to recommend the airline by State and Airport City (R-Studio)





Boxplot of Customer Loyalty (R-Studio)

Barplot of Customer Loyalty (R-Studio)

Conclusion and Learning Goals:

Basic data analysis is the backbone of every data analyst and scientist working in the field nowadays and this course is what I attribute to being that start for me. Not only did I begin to feel the strange sense of satisfaction that comes forth when the code you write gives you the answers you're looking for (be it a visualization or regression model), but the course and project also allowed me to showcase some of those basic skills I knew I would be utilizing going forward. This project allowed me to begin thinking critically when it comes to data as well, giving me the opportunity to find solutions to problems that may arise in any type of dataset I work with in the future.

This project and class overall contributed to the application of some of the learning goals set forth by the program such as:

- Utilizing my skills to describe a broad overview of many basic data science areas of practice learned in the class.
- Identifying patterns through visualization, classification, and regression models I created.
- Developing alternative strategies and a plan of action based on the data and its analysis.
- Demonstrating my communications skills of the statistical analysis in a way that could be easily interpreted by all.

Syracuse University School of Information Studies
M.S. Applied Data Science

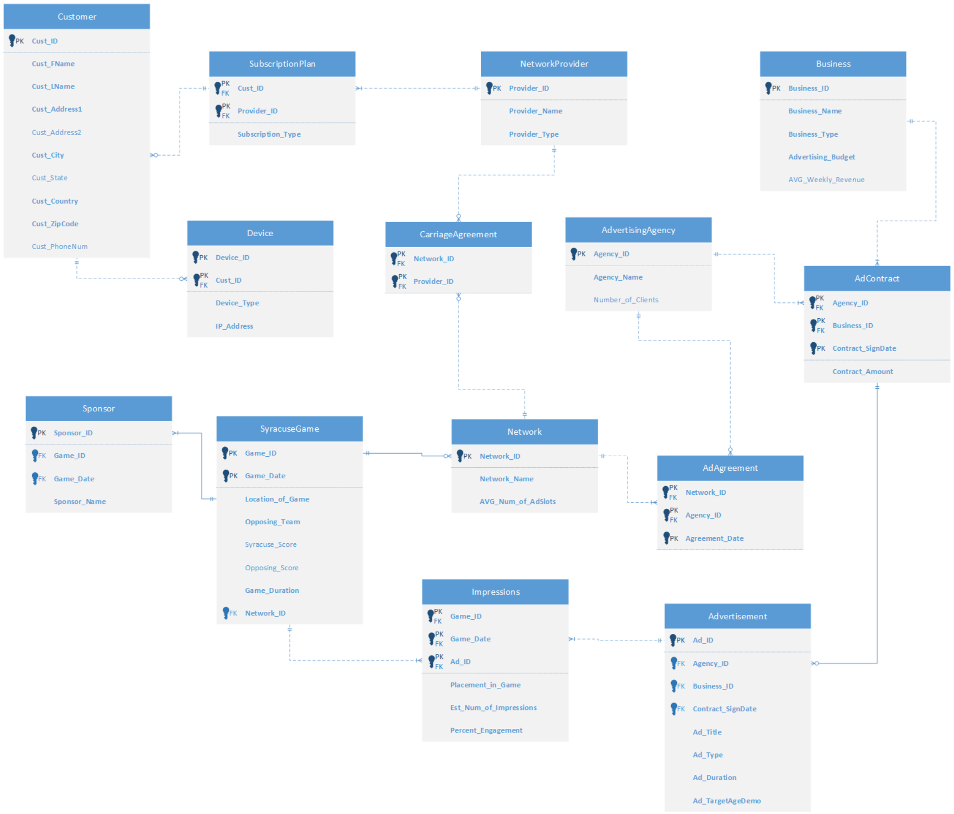
IST 659 - Data Administration Concepts and Database Management

Project Overview:

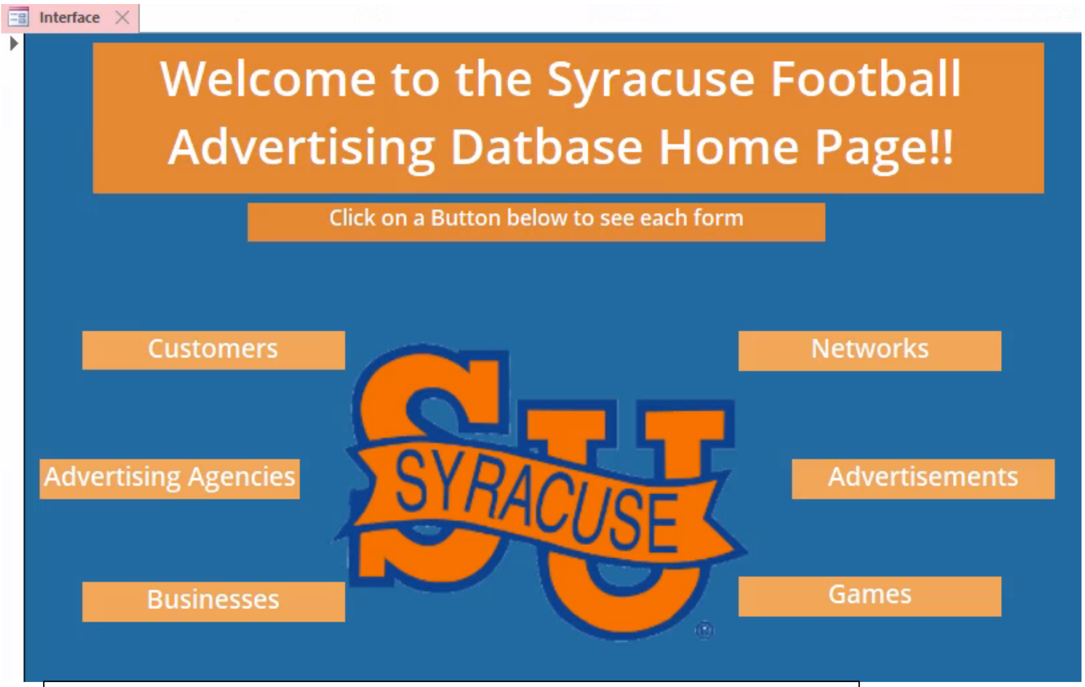
Under the direction of Professor Hernando Hoyos, I was a part of a group that created an Advertising Database that could be utilized during Syracuse University's Football season that was taking place at the time. Within the database, my partner and I tried to focus our attention on the businesses doing the advertising, the streaming platforms that might be showing the games, the games themselves, different local and national advertising agencies, and the individual users who would be most likely to watch Syracuse Football. Because of the many different types of data we would need to implement our system overall, we set out to both collect, create, and organize the data that we saw fit.

Models were developed to organize the relationships between consumers, streaming services, advertisers, local and national companies, television networks, and Syracuse Athletics utilizing the skills we learned through Visio and SQL Server Management Studio. Data population was then accomplished using Microsoft Access, where we also created the Forms and Report that would be implemented to input and store new data, as well as report the results after a game. Basic forms were created form consumers, agencies, businesses, and individual advertisements, while reports were created to showcase the ads that received the most impressions, the type of devices that received the most use, and the most popular type of subscription plan per streaming service.

Screenshot Examples from Project:



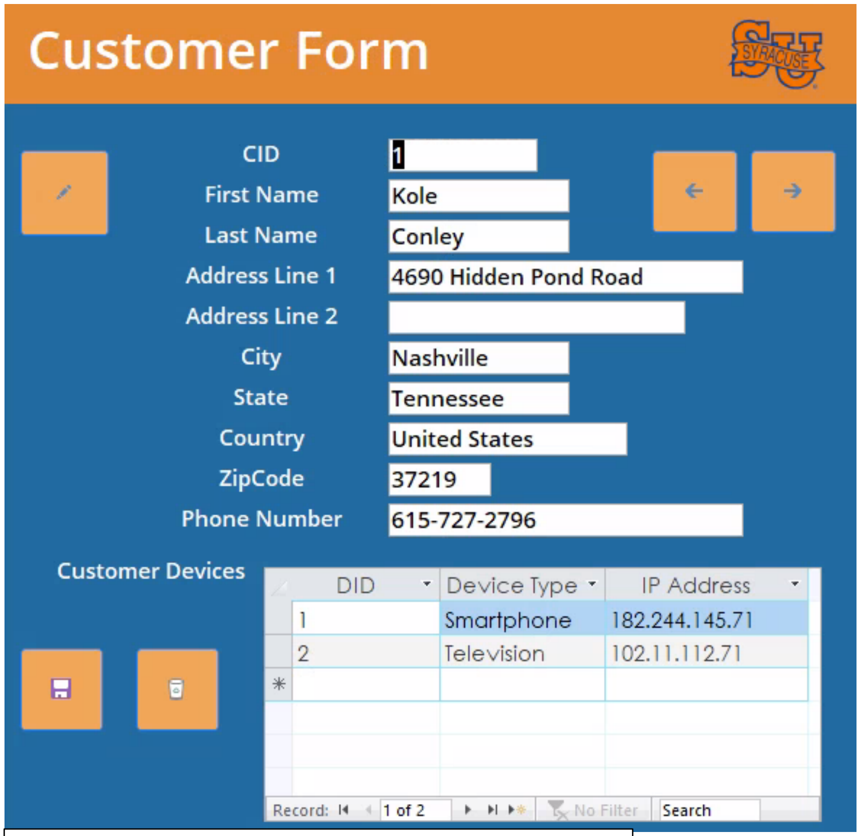
Basic model of relationships between tables in the system (Visio)



Welcome to the Syracuse Football Advertising Datbase Home Page!!

Click on a Button below to see each form

Hub to access forms made for Database (Microsoft Access)



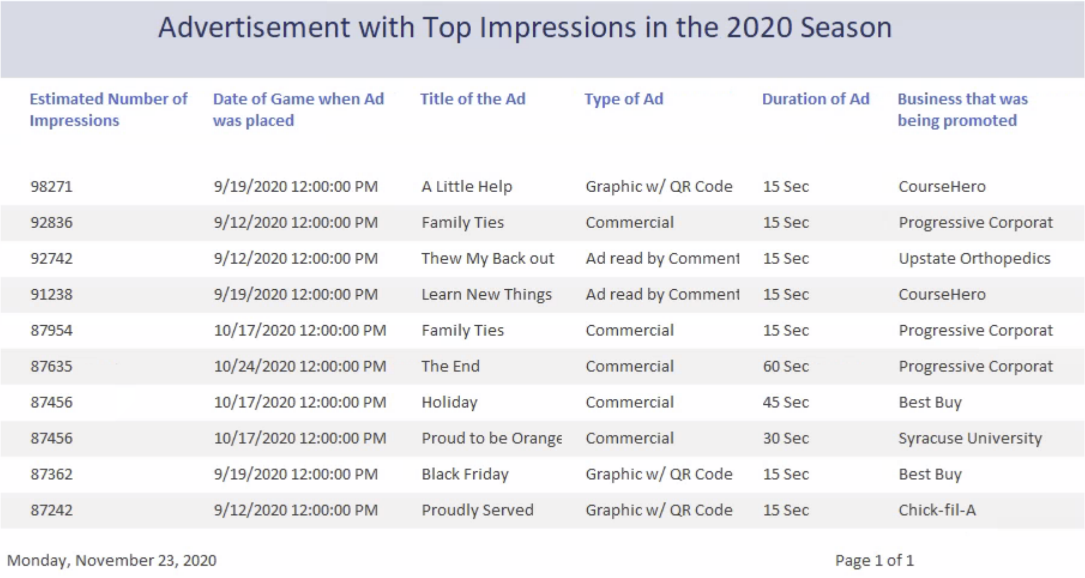
Customer Form

CID	1
First Name	Kole
Last Name	Conley
Address Line 1	4690 Hidden Pond Road
Address Line 2	
City	Nashville
State	Tennessee
Country	United States
ZipCode	37219
Phone Number	615-727-2796

Customer Devices

DID	Device Type	IP Address
1	Smartphone	182.244.145.71
2	Television	102.11.112.71

Example Customer Form (Microsoft Access)



Advertisement with Top Impressions in the 2020 Season

Estimated Number of Impressions	Date of Game when Ad was placed	Title of the Ad	Type of Ad	Duration of Ad	Business that was being promoted
98271	9/19/2020 12:00:00 PM	A Little Help	Graphic w/ QR Code	15 Sec	CourseHero
92836	9/12/2020 12:00:00 PM	Family Ties	Commercial	15 Sec	Progressive Corporat
92742	9/12/2020 12:00:00 PM	Thew My Back out	Ad read by Comment	15 Sec	Upstate Orthopedics
91238	9/19/2020 12:00:00 PM	Learn New Things	Ad read by Comment	15 Sec	CourseHero
87954	10/17/2020 12:00:00 PM	Family Ties	Commercial	15 Sec	Progressive Corporat
87635	10/24/2020 12:00:00 PM	The End	Commercial	60 Sec	Progressive Corporat
87456	10/17/2020 12:00:00 PM	Holiday	Commercial	45 Sec	Best Buy
87456	10/17/2020 12:00:00 PM	Proud to be Orange	Commercial	30 Sec	Syracuse University
87362	9/19/2020 12:00:00 PM	Black Friday	Graphic w/ QR Code	15 Sec	Best Buy
87242	9/12/2020 12:00:00 PM	Proudly Served	Graphic w/ QR Code	15 Sec	Chick-fil-A

Monday, November 23, 2020 Page 1 of 1

Example Advertisement Report (Microsoft Access)

Conclusion and Learning Goals:

The creation and management of data within a warehouse setting in order to find solutions to problems that may arise through the data, as I've said before, is extremely important in the world of data science today. Having the ability to work through this process from proposal to implementation and analysis helped me to understand the importance of how data is stored and accessed. The insights that my partner and I were able to come up with through the basic analysis we did in Microsoft Access allowed us to continuously critically think about the business rules and problems that we were dealing with, along with how we might be able to meet those rules/fix those issues in the long run. We also made a conscious attempt through our business rules to ensure that, if our basic program were to ever be actually implemented, the privacy of the individuals we have in the system would not have their personal information used in ways they may not want it to.

This project and class overall contributed to the application of the learning goals set forth by the program such as:

- Collecting/creating data and managing that data for implementation into our system.
- Identifying patterns through the forms and reports we created.
- Developing alternative strategies and a plan of action based on the results we "received".
- Demonstrating our communications skills of the statistical analysis in a way that could be easily interpreted by all.
- Dealing with ethical dilemmas that may arise when working with certain types of data.

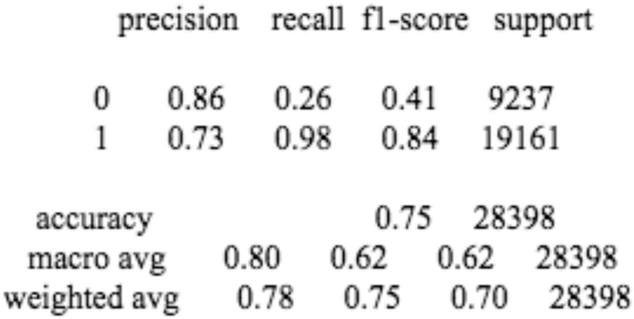
IST 718 – Big Data Analytics

Project Overview:

Under the direction of Professor Daniel Acuna, my team members and I were able to build our knowledge of regression, classification, and clustering even further than we did in previous classes by utilizing the methods taught throughout the course in Python and PySpark, with a heavy emphasis on machine learning. Our project dealt with data from school districts across the United States to see if there was any connection between school spending and the overall performance of their students. Focusing on whether the districts passed the benchmark we as a group decided on, we were then able to formulate suggested solutions to create what we would refer to as "equal opportunities" for all students.

Data was collected from various sources, which was then pieced together and cleansed using R-Studio. Feature engineering was also done in R-Studio as a part of the dataset's preparation. Logistic regression, random forest classification, and k-means clustering was then performed within the JupyterHub server we used throughout the semester. From the creation of training, validation, and testing datasets, we were able to tune the parameters of each individual model to find the one that most accurately helped to predict the variable we set out to determine, leading us to find the variables in the dataset that had the most impact on these school districts. Finally, we were able to come up with solutions for school districts in order to be able to meet the academic standards that would allow each and every single student to thrive to the best of their ability.

Screenshot Examples from Project:

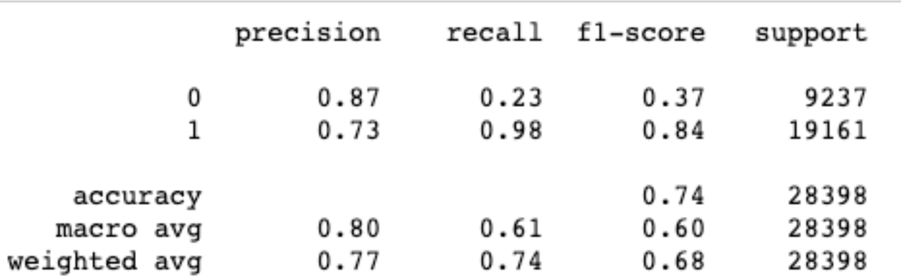


```
              precision    recall  f1-score   support

           0       0.86      0.26      0.41      9237
           1       0.73      0.98      0.84     19161

    accuracy                           0.75     28398
   macro avg       0.80      0.62      0.62     28398
weighted avg       0.78      0.75      0.70     28398
```

Accuracy Report from Random Forest (JupyterHub/PySpark)

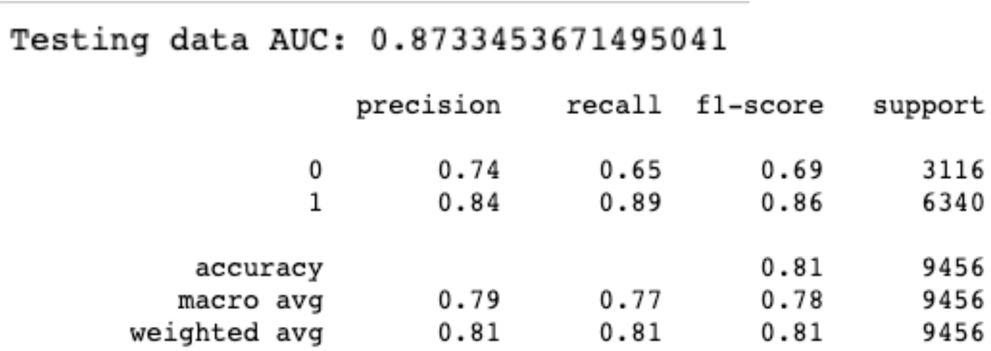


```
              precision    recall  f1-score   support

           0       0.87      0.23      0.37      9237
           1       0.73      0.98      0.84     19161

    accuracy                           0.74     28398
   macro avg       0.80      0.61      0.60     28398
weighted avg       0.77      0.74      0.68     28398
```

Accuracy Report from Lasso Regression (JupyterHub/PySpark)



```
Testing data AUC: 0.8733453671495041
              precision    recall  f1-score   support

           0       0.74      0.65      0.69      3116
           1       0.84      0.89      0.86      6340

    accuracy                           0.81      9456
   macro avg       0.79      0.77      0.78      9456
weighted avg       0.81      0.81      0.81      9456
```

Area Under the Curve and Accuracy Report from Best Model (JupyterHub/PySpark)

Conclusion and Learning Goals:

This course, overall, had to have been both the hardest and most rewarding class I took during the entire program. As someone who, going into the class, wasn't a very knowledgeable person when it comes machine learning, I was able to use this class and Professor Acuna's teachings to give me a base understanding for the subject going forward, especially when it comes to languages like Python and Pandas (to an extent). By the end of the course, some of the coding when it comes to the creation of regression and classification models became second nature. Professor Acuna, along with my groupmates, also helped me develop communications skills when it comes to data science, allowing me to not only interpret the results of my analysis but to express my findings to other members of the data science field as well.

This project and class overall contributed to the application of the learning goals set forth by the program such as:

- Collecting and managing/cleansing data to be used to draw our conclusions through analysis.
- Identifying patterns through the analysis and clustering methods we decided to use.
- Developing alternative strategies and a plan of action based on the results we found.
- Demonstrating our communications skills of the statistical analysis in a way that could be easily interpreted by all.
- Dealing with ethical dilemmas that may arise when working with certain types of data.

Syracuse University School of Information Studies
M.S. Applied Data Science

IST 719 – Information Visualization

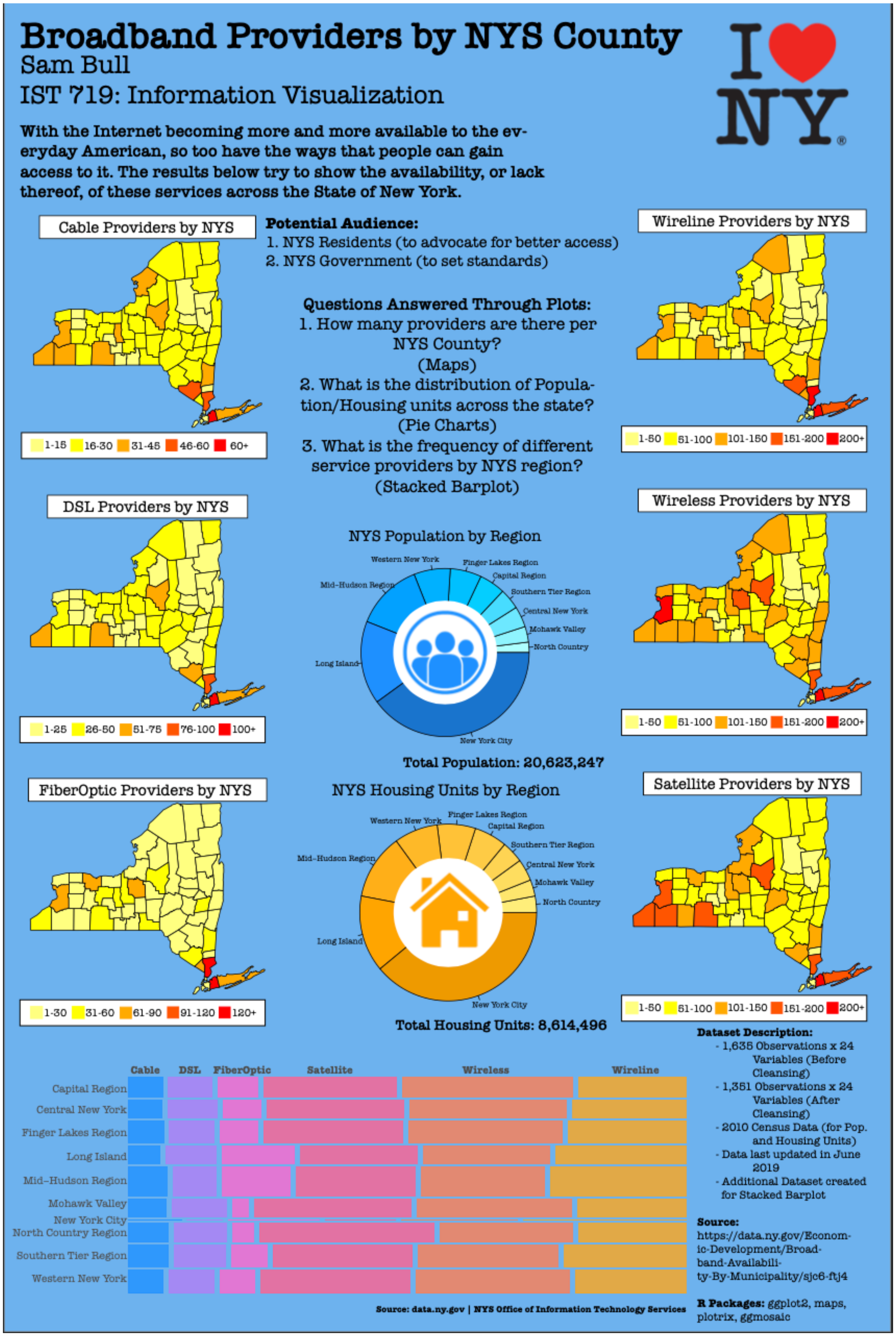
Project Overview:

Under the direction of Professor Jeffery Hemsley, our class was tasked with finding a dataset that could be utilized to create interesting, creative, and interpretable visualizations for the final poster session at the end of the semester. For my final project, I decided to try and find any disparities between counties in the state of New York when it came to their ability to connect with different types of broadband service providers. Throughout the course of the semester, we discussed numerous aspects for the creation of both different types of charts, graphs, and maps as well as for creating an appealing poster in order to get a message across to any and all onlookers. Utilizing R-Studio, with packages like ggmoasic and colorbrewer, and Adobe Illustrator, I gained the skills to create complex and visually pleasing ways of helping to answer any question that may arise from any given dataset.

Getting the data itself from data.ny.gov, I was able to cleanse and separate the variables I wanted to focus on that would, hopefully, help me find any disparities. Once the data itself was set, and the R-packages I needed were added to my workspace, I utilized a number of the different techniques and skills I learned throughout the semester to create basic distribution and analysis plots that I wanted to include to answer any questions I set out to answer at the beginning of the project. From there, I imported these plots into Adobe Illustrator, where I was able to change them in such aesthetic ways to be more appealing once they would be added to the final poster. The poster itself was then created and changed in Illustrator as well. Finally, the presentation of the final product allowed me to not only discuss my thought process with the work that I did, but also give my thoughts on any problems that I found while going through the data, as well as some solutions I had in regard to those problems.

Syracuse University School of Information Studies
M.S. Applied Data Science

Screenshot from the Project:



Broadband Providers by NYS County

Sam Bull

IST 719: Information Visualization

With the Internet becoming more and more available to the everyday American, so too have the ways that people can gain access to it. The results below try to show the availability, or lack thereof, of these services across the State of New York.

Potential Audience:
1. NYS Residents (to advocate for better access)
2. NYS Government (to set standards)

Questions Answered Through Plots:
1. How many providers are there per NYS County? (Maps)
2. What is the distribution of Population/Housing units across the state? (Pie Charts)
3. What is the frequency of different service providers by NYS region? (Stacked Barplot)

Cable Providers by NYS

DSL Providers by NYS

FiberOptic Providers by NYS

Wireline Providers by NYS

Wireless Providers by NYS

Satellite Providers by NYS

NYS Population by Region

Total Population: 20,623,247

NYS Housing Units by Region

Total Housing Units: 8,614,496

Dataset Description:
- 1,635 Observations x 24 Variables (Before Cleansing)
- 1,351 Observations x 24 Variables (After Cleansing)
- 2010 Census Data (for Pop. and Housing Units)
- Data last updated in June 2019
- Additional Dataset created for Stacked Barplot

Source: https://data.ny.gov/Economic-Development/Broadband-Availability-By-Municipality/sjc6-ftj4

R Packages: ggplot2, maps, plotrix, ggmosaic

Source: data.ny.gov | NYS Office of Information Technology Services

Final Poster (R-Studio/Adobe Illustrator)

Conclusion and Learning Goals:

Information visualization as a course allowed me to finally gain skills to build upon many of the different plots I learned to create at the start of the program in July of 2020. Data visualizations help people to understand any message or statement you might try to make through your analysis and the ability to create elaborate and eye-catching images can be extremely important. Through the collection and management of the data, its analysis, and its final presentation through the poster, I was able to take a number of the skills that I learned not only in this class but from my work across this program to create something that, I believe, helped to answer the questions I wanted to find answers to.

This project and class overall contributed to the application of the learning goals set forth by the program such as:

- Collecting and managing/cleansing data to be used for the visuals.
- Identifying patterns through the analysis and clustering methods we decided to use.
- Demonstrating my communications skills of the statistical analysis in a way that could be easily interpreted by all.

Overall Conclusion:

This portfolio demonstrates how the different skills and techniques in the world of data science that were taught throughout the different courses in the ADS program fulfill the seven learning objectives dealing with the major areas of data science. For a number of the projects, data was obtained through different sources, be it from the instructor or from my own searching in databases, managed through programs such as R-Studio and JupyterHub, and were analyzed through the different regression, classification, and clustering methods taught, as well as implementation of alternative strategies in order to find the one(s) that fit the data the best. Several visualizations were created from that data analysis in order to both better demonstrate the questions being answered as well as show the scale and scope of the data itself for the questions to arise. From both the analysis and the visualizations, actionable recommendations could be made to help push the work being done from just data analysis to actual change that could be seen in the world, such as the suggestions made based on the feature importance of certain variables determining if a school's students were successful or not.

My communication skills improved dramatically as well and were displayed through the expression of the insights and solutions found through the analysis, classification, and clustering done across the projects and their many presentations. The ethical dimensions of the data science practice were also exemplified through both the collection and creation of data. By selecting certain variables within each dataset with the intent of only utilizing those that would aid in the analysis process, I attempted to mitigate the dilemmas that could arise when working with personal identification information, such as with school districts across America or with streaming service users.

The School of Information Studies at Syracuse University allow students to learn the skills necessary to meet the different learning goals they set for those students. Through the collection and management of data, along with the implementation of different analysis strategies and visualization techniques, useful insights and suggestions could be made to fix any problem that may present itself. Because I was a part of this program, I know feel as though I have the skills, both technically and ethically, to tackle any number of a wide range of data analysis problems that may find me going forward in my career.