Syracuse University

Applied Data Science Summer 2021 Portfolio Presentation

Samuel Bull
SUID#: 984005814
stbull@syr.edu

Presentation Overview

Introduction

- **Purpose of Presentation/Learning Objectives to meet**

Applications

- **IST 687 - Introduction to Data Science**
  - **Course/Project Descriptions with Examples**
  - **Conclusions/Learning Goals met**
- **IST 659 - Data Administration Concepts and Database Management**
  - **Course/Project Descriptions with Examples**
  - **Conclusions/Learning Goals met**
- **IST 718 - Big Data Analytics**
  - **Course/Project Descriptions with Examples**
  - **Conclusions/Learning Goals met**
- **IST 719 - Information Visualization**
  - **Course/Project Descriptions with Examples**
  - **Conclusions/Learning Goals met**

Overall Conclusion

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.
- 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.
- 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:
    - Describe a broad overview of the major practice areas in data science.
    - Collect and organize data
    - Identify Patterns in data via visualization, statistical analysis, and data mining.
    - Develop alternative strategies based on the data.
    - Develop a plan of action to implement the business decisions derived from the analyses.
    - Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization.
    - Synthesize the ethical dimensions of data science practice (e.g., privacy)

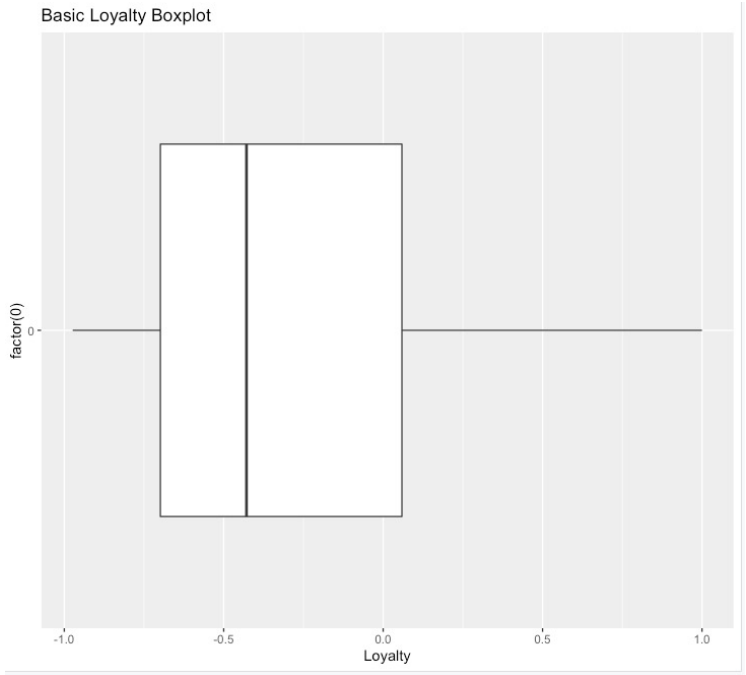
IST 687 - Introduction to Data Science

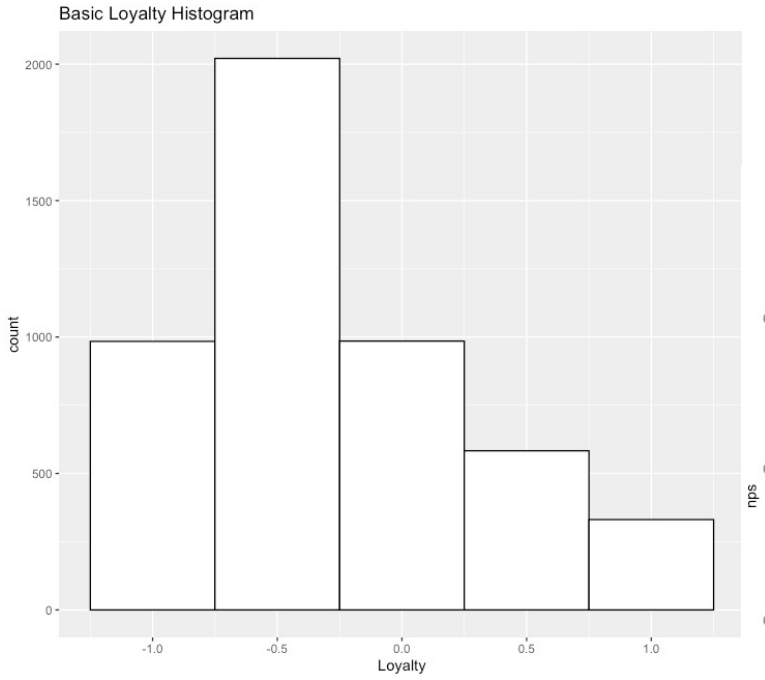
Professor Jeffery Saltz

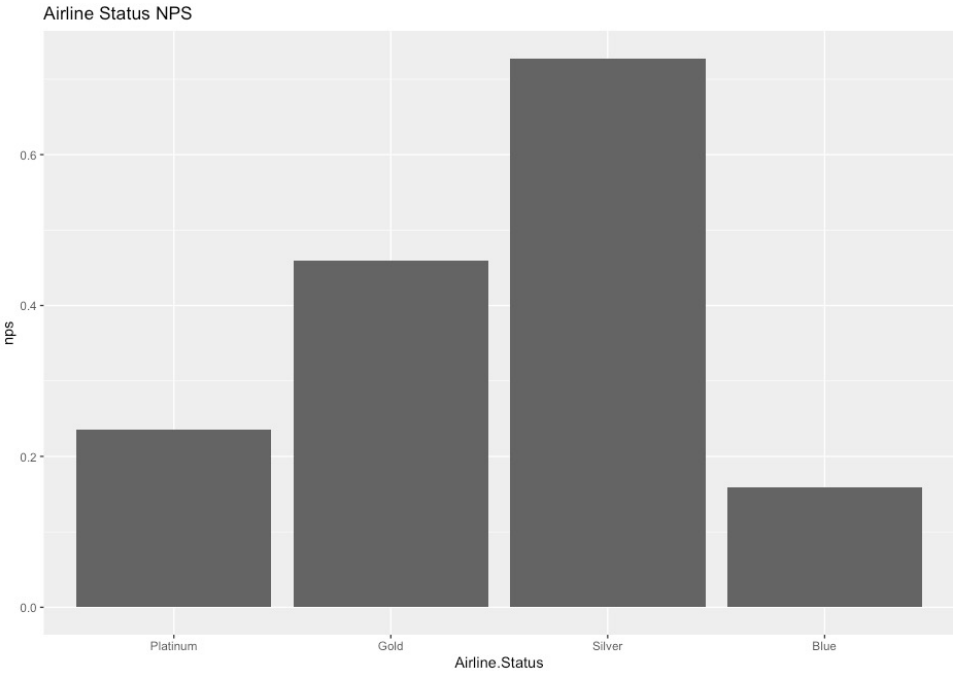
Course/Project Description

- This course is what I consider to be my introduction to R and R-Studio
- Professor Saltz taught everything from the most basic mathematical commands to different analysis/classification methods and both basic and advanced visualizations (ggplot2, maps, etc.)
- Project consisted of cleansing and working with data from “Southeast Airlines” to determine whether or not frequent fliers would recommend the airline to their friends
- Work was done utilizing boxplot, bar charts, histograms, maps, association rules mining, linear model regression, and Support Vector Machines
- A PowerPoint Presentation was then created to display my findings and offer solutions

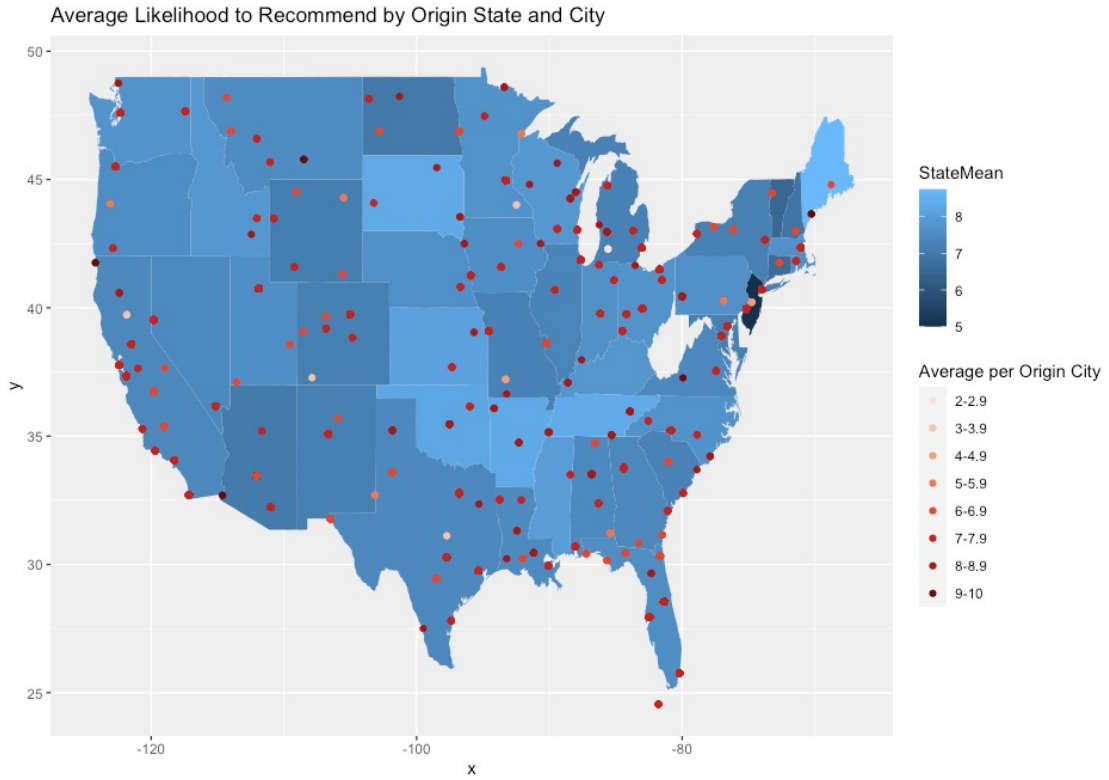
Example of Charts from Project

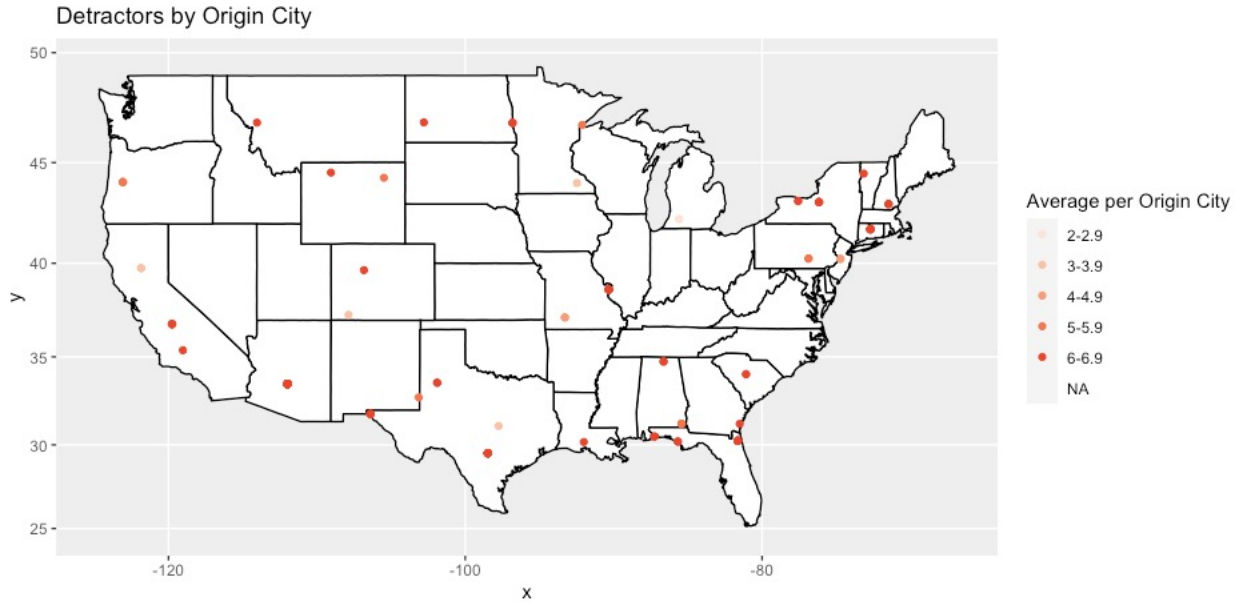






Example of Charts from Project



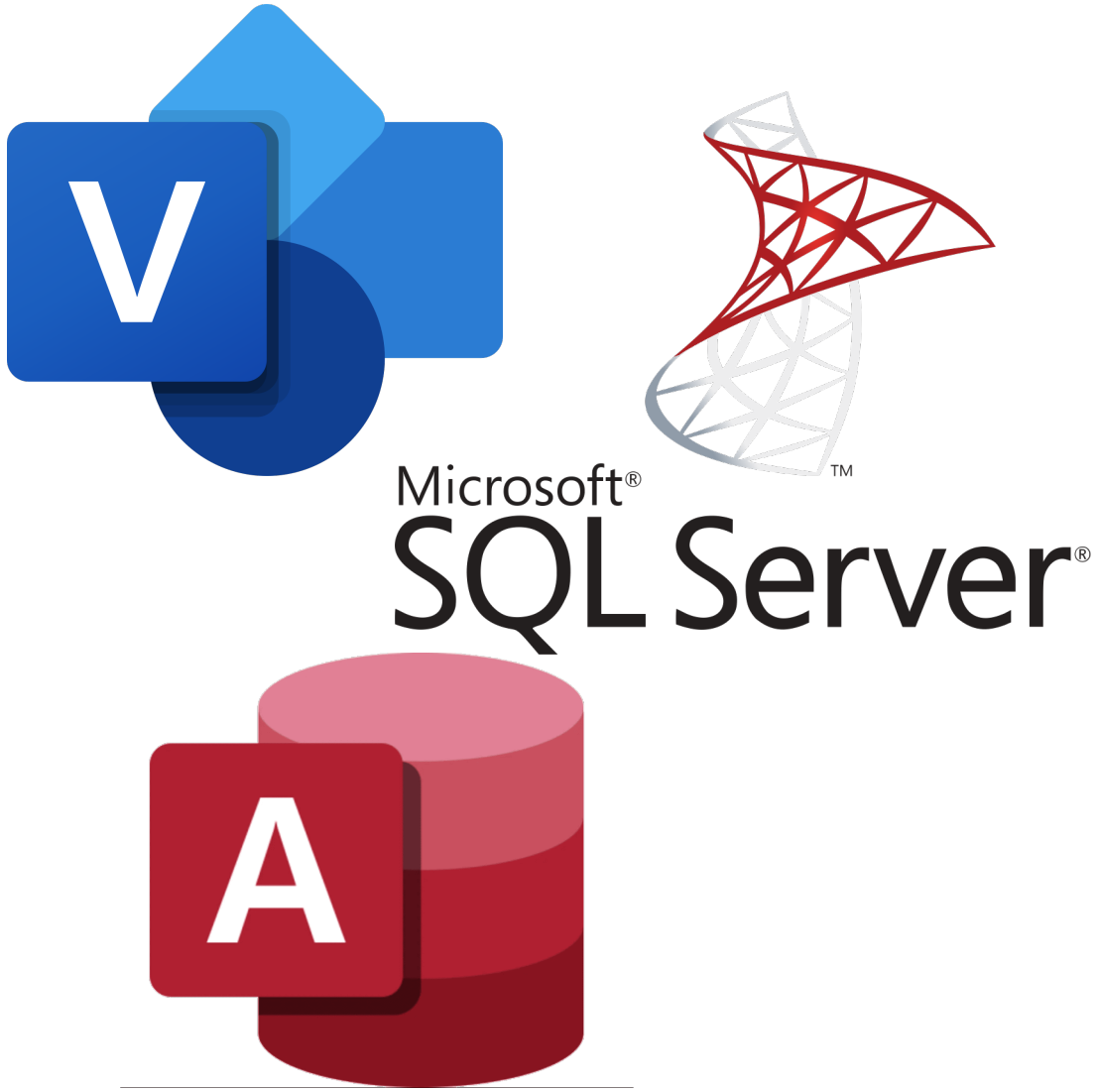


Conclusions/Learning Goals

- The learning goals I met through this project and the ways I went about meeting them are:
  - Utilizing my skills to describe a broad overview of many basic data science areas of practice learned in the class
    - Working with visualizations, classification, and regression methods and models
  - Identifying patterns through visualization, classification, and regression models I created
    - Comparing charts to one another, changing parameters within the SVM and linear regression models
  - Developing alternative strategies and a plan of action based on the data and its analysis
    - Creating suggestions for Southeast Airlines to help bring about more positive feedback from frequent fliers
  - Demonstrating my communications skills of the statistical analysis in a way that could be easily interpreted by all
    - Presenting my findings once all was said and done to my professor

IST 659 - Data Administration Concepts and Database Management

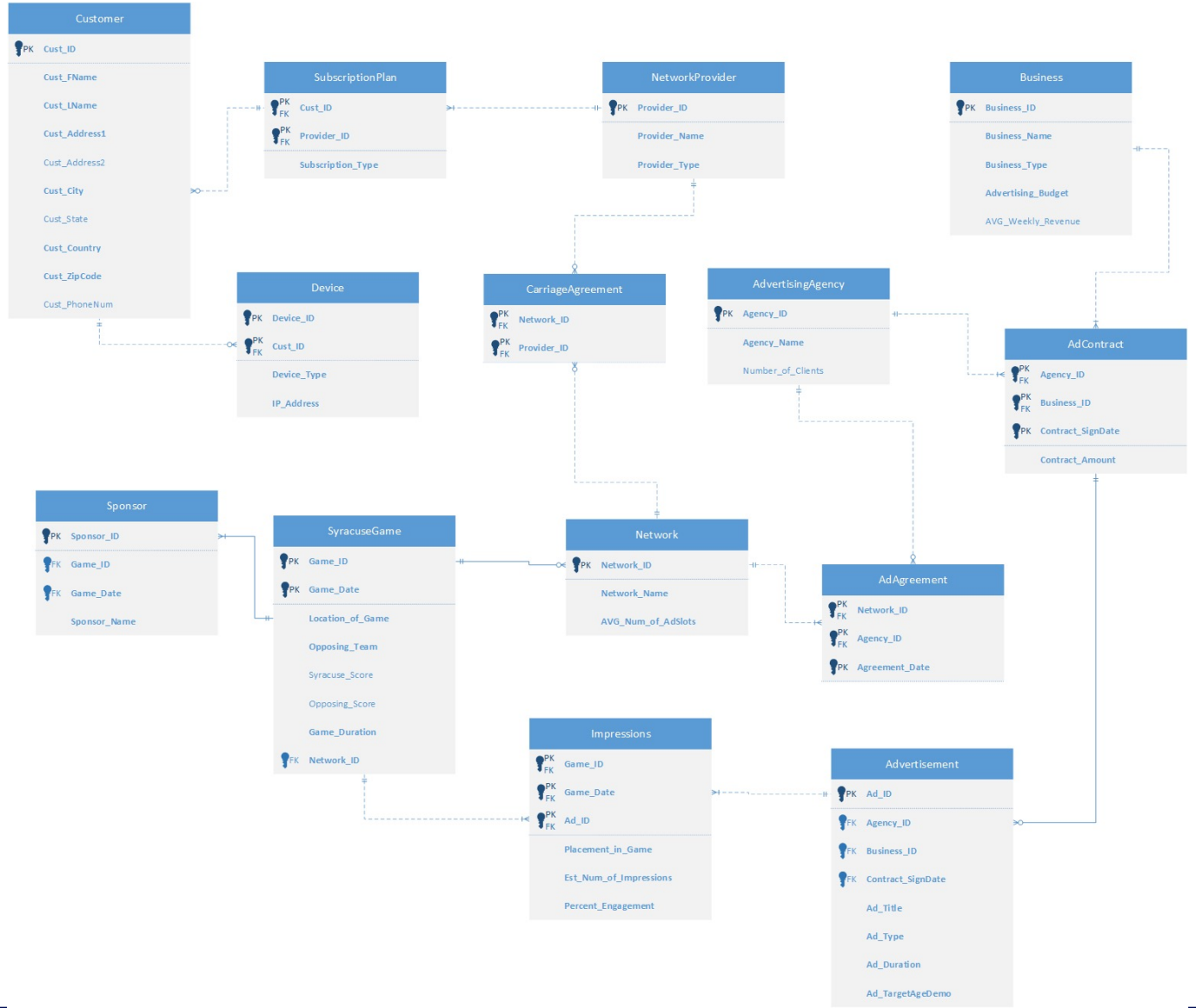
Professor Hernando Hoyos



Course/Project Description

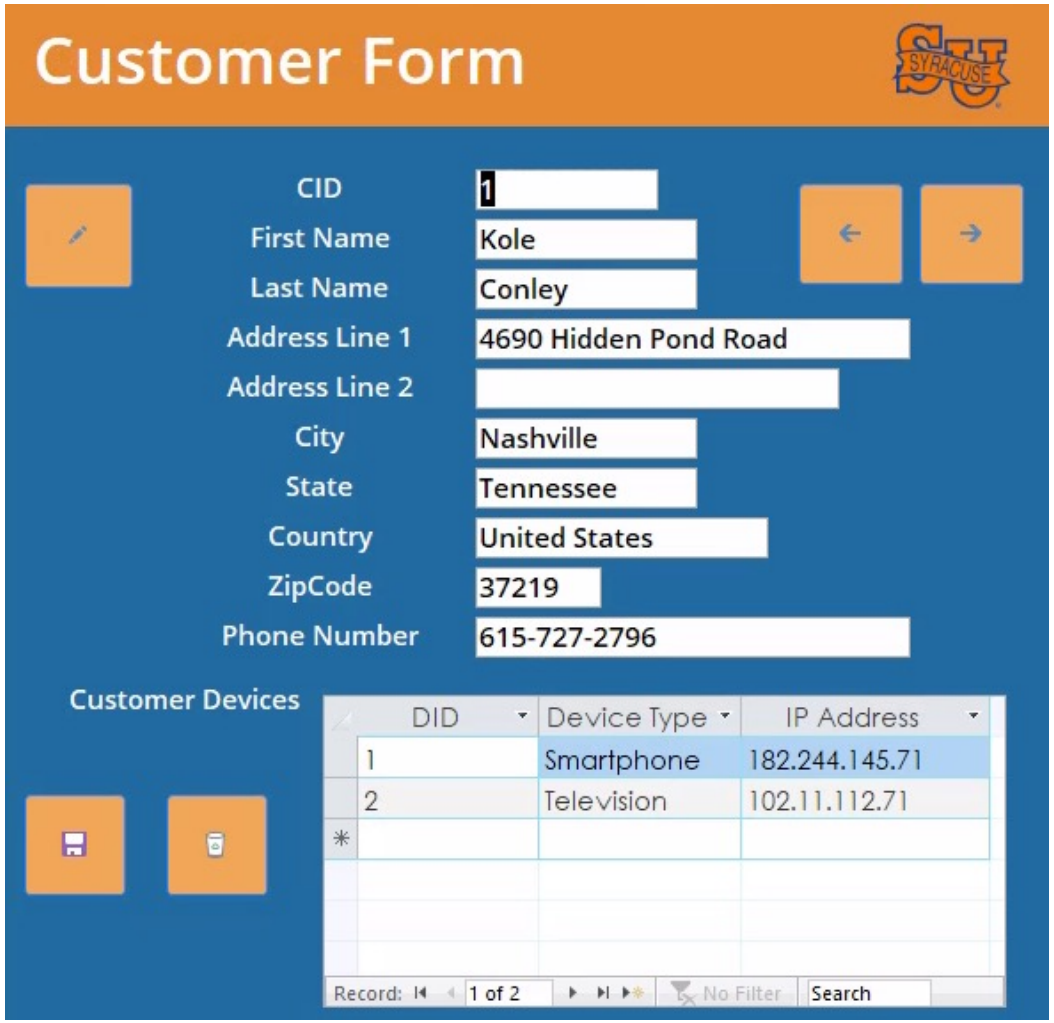
- This course allowed me to build upon SQL and database skills I started to learn during my undergraduate collegiate career
- Professor Hoyos taught methods of building both basic and complex database management systems through Microsoft Visio, SQL Server, and Access
  - Along with the creation of forms and reports
- My partner and I created a system where companies can see how their advertisements affect consumers during Syracuse Football games being streamed online
- Tables were created, relationships were established, data was both collected and created, and forms/reports were created from the data
- A PowerPoint Presentation was created and presented in front of the class

Example of Relationships from Project



Example of Layout from Project

Example of Form/Report from Project



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

Conclusions/Learning Goals

- **The learning goals I met through this project and the ways I went about meeting them are:**
  - **Collecting/creating data and managing that data for implementation into our system.**
    - **Collection of local and national advertiser information, streaming platforms, and company names while creating information about users, their devices, and specific ad spots**
  - **Identifying patterns through the forms and reports we created.**
    - **Answering business questions we set out to answer via the forms and, much more specifically, the reports we as a collective created**
  - **Developing alternative strategies and a plan of action based on the results we "received".**
    - **Suggesting what companies/advertisers should do to pull in more media attention based on consumers**
  - **Demonstrating our communications skills of the statistical analysis in a way that could be easily interpreted by all.**
    - **Presenting our findings/experience for both our professor and our peers**
  - **Dealing with ethical dilemmas that may arise when working with certain types of data.**
    - **Setting business rules in order to protect the privacy of the "consumers" within our system**

IST 718 - Big Data Analytics

Professor Daniel Acuña

Course/Project Description

- This course was both and incredibly difficult and rewarding experience
- Professor Acuña taught concepts in machine learning and data analysis while working in PySpark and Pandas within a Jupyter Notebook Server
- My group and I collected data from numerous sources to determine whether budget spending in US school districts helps the success rate of its students
- Data was cleansed in R-Studio and the Random Forest/Lasso(Logistic) Regression and K-Means Clustering was done in PySpark
- Once we got our results/determined the better way of predicting, we created a PowerPoint Presentation to display our findings and offer solutions for school districts

Example of Analysis/Tables from Project

Random Forest

	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

Lasso(Logistic) Regression

	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

```
Logistic Regression Model AUC: 0.8792980250308499
```

Example of Charts from Project

Best Model (Lasso)

	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

Testing data AUC: 0.8733453671495041

Conclusions/Learning Goals

- **The learning goals I met through this project and the ways I went about meeting them are:**
  - Collecting and managing/cleansing data to be used to draw our conclusions through analysis.
    - Finding Census data about counties within the United States, searching for and modifying school district information, and combining different datasets into one to utilize through our analysis
  - Identifying patterns through the analysis and clustering methods we decided to use.
    - Through analysis and clustering, we found patterns within certain budget elements that would either help or hurt a school districts success rate
  - Developing alternative strategies and a plan of action based on the results we found.
    - From the items we found to help/hurt schools, we gave hypothetical solutions in order to create equal opportunity
  - Demonstrating our communications skills of the statistical analysis in a way that could be easily interpreted by all.
    - Presenting updates as well as our final findings to our professor and class throughout the semester
  - Dealing with ethical dilemmas that may arise when working with certain types of data.
    - Working around problems that came with certain information dealing with specific details of school districts

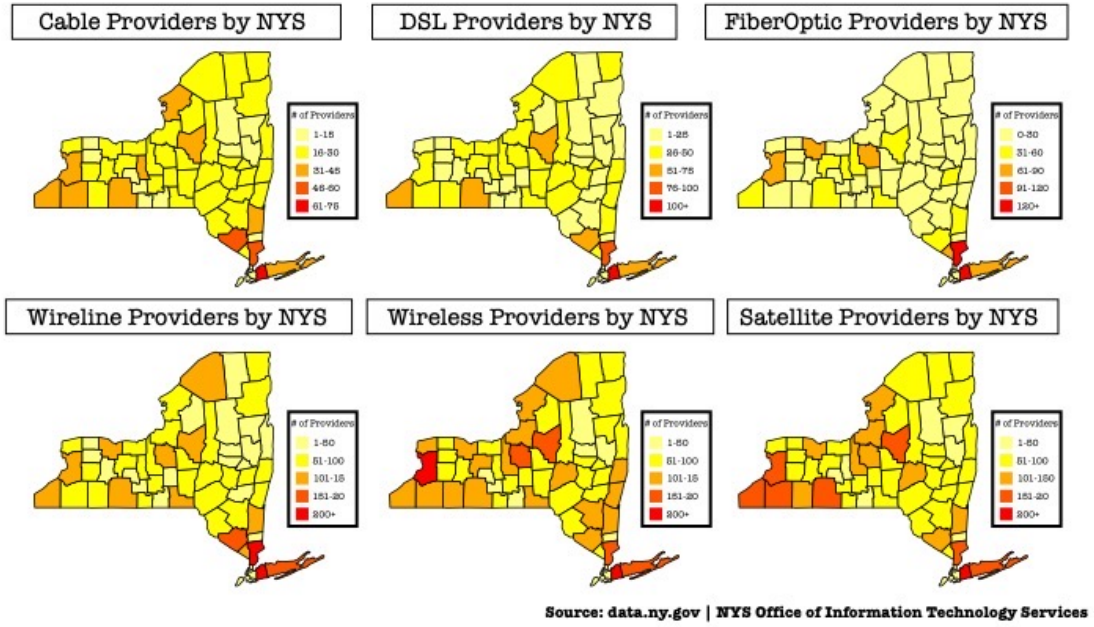
IST 719 - Data Visualization

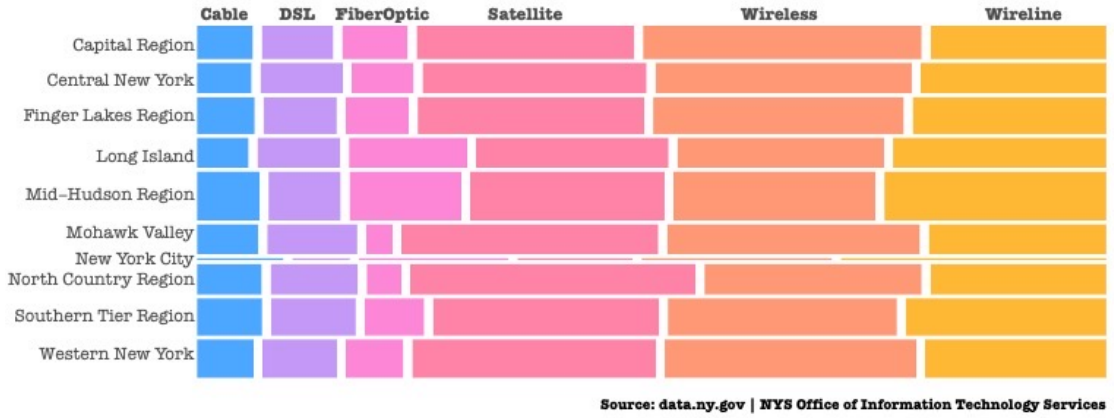
Professor Jeffery Hemsely

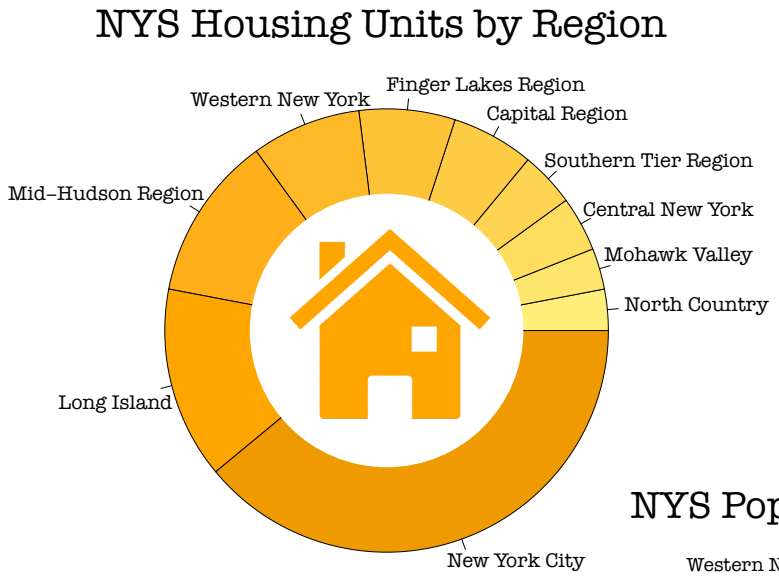
Course/Project Description

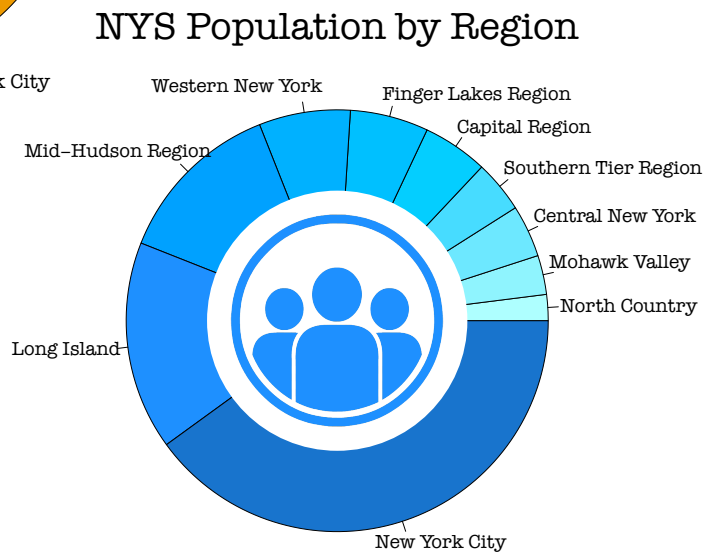
- This course helped to build upon the basic visualization and editing skills I learned in IST 687
- Professor Hemsley taught design and layout principles/skills to create different visualizations in R-Studio, import them into Adobe Illustrator, and make them more dynamic/aesthetically pleasing
- I collected data from data.ny.gov via the NYS Office of Information Technology Services dealing with the distribution of different Broadband connectivity methods in across the state
- Utilizing packages in R-Studio (ggplot2, ggmosaic, colorbrewer, etc.), I created basic plots and then made them more dynamic in Illustrator
- The final poster was also created in Illustrator, which was presented in front of my class at the semesters end

Example of Charts from Project

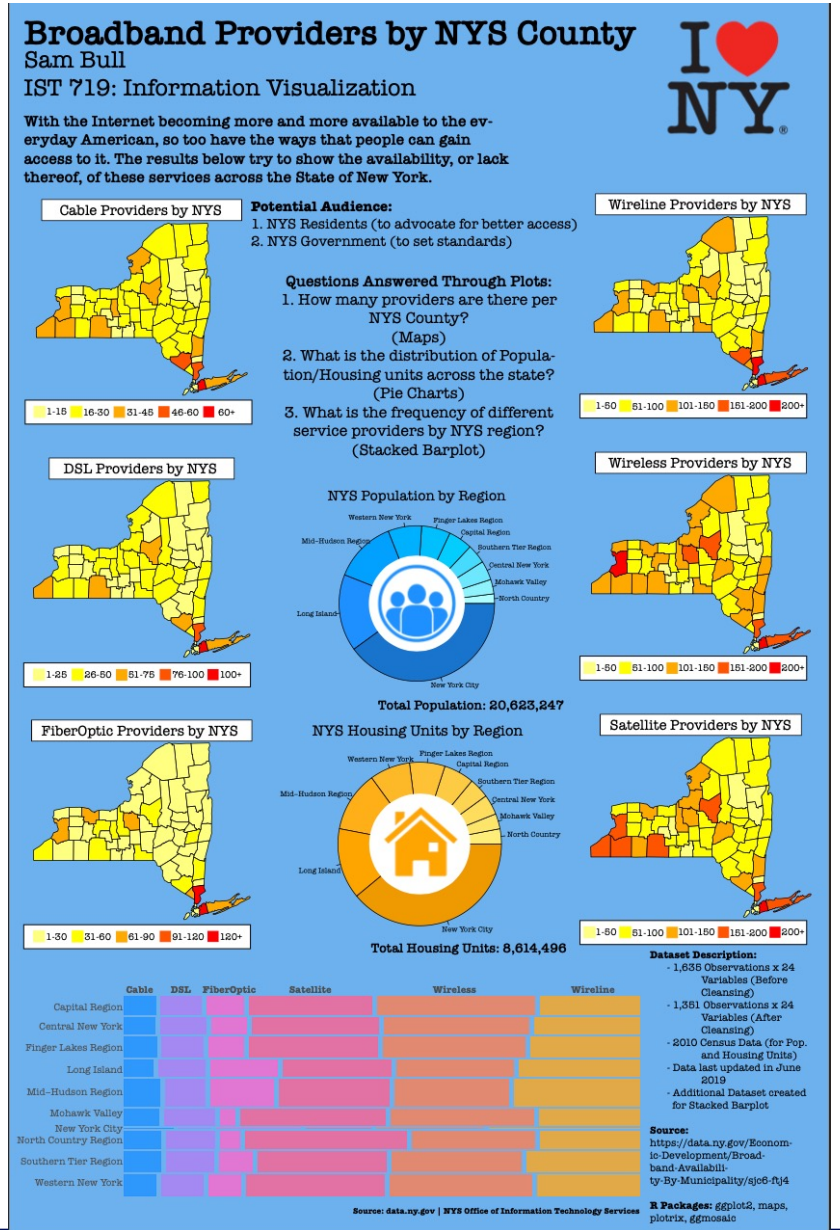








Final Poster



Conclusions/Learning Goals

- **The learning goals I met through this project and the ways I went about meeting them are:**
  - **Collecting and managing/cleansing data to be used for the visuals.**
    - **Finding data on data.ny.gov, cleaning it to only utilize the variables I needed to use, taking out variables that contained data that would hinder some of the visualizations, and create a separate dataframe entirely for a certain visualization**
  - **Identifying patterns through the analysis and clustering methods we decided to use.**
    - **Utilizing distribution charts to see where improvements within the state of New York were needed for certain broadband provider methods**
  - **Demonstrating my communications skills of the statistical analysis in a way that could be easily interpreted by all.**
    - **Presenting the poster in a way as to get my message across to both my professor, the TA, and my peers**

Overall Conclusions

- 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 while all attempting to describe both broad and more specific areas of data science.
  - For a number of the projects, data was obtained through different sources (IST 652, IST 718, IST 719)
  - Data was analyzed through the different regression, classification, and clustering methods taught (IST 687, IST 652, IST 718, IST 719)
  - The implementation of alternative strategies in order to find the one(s) that fit the data the best (IST 687, IST 718)
  - Actionable recommendations were made to help solve problems that were being posed by the data as well as myself (IST 687, IST 652, IST 718)
  - My communication skills improved dramatically as well and were displayed through the expression of the insights and solutions found through analysis (IST 687, IST 652, IST 718, IST 719)
  - The ethical dimensions of data science practices were also exemplified through both the collection and creation of data (IST 652, IST 718)

Syracuse University

Thank You!

Samuel Bull
Masters Student in Applied Data Science
SUID#: 984005814
stbull@syr.edu