Syracuse University

Master of Science
Applied Data Science

Portfolio Milestone

Kyle Umlang

SUID 109166863
March 19, 2021
Link to Portfolio Folder

Introduction

The Applied Data Science program at Syracuse University's School of Information Studies prepares students for a career in data science by teaching them how to analyze and operationalize all sizes of data sets, understand information science and management principles and applying data science to enterprise operations and processes.

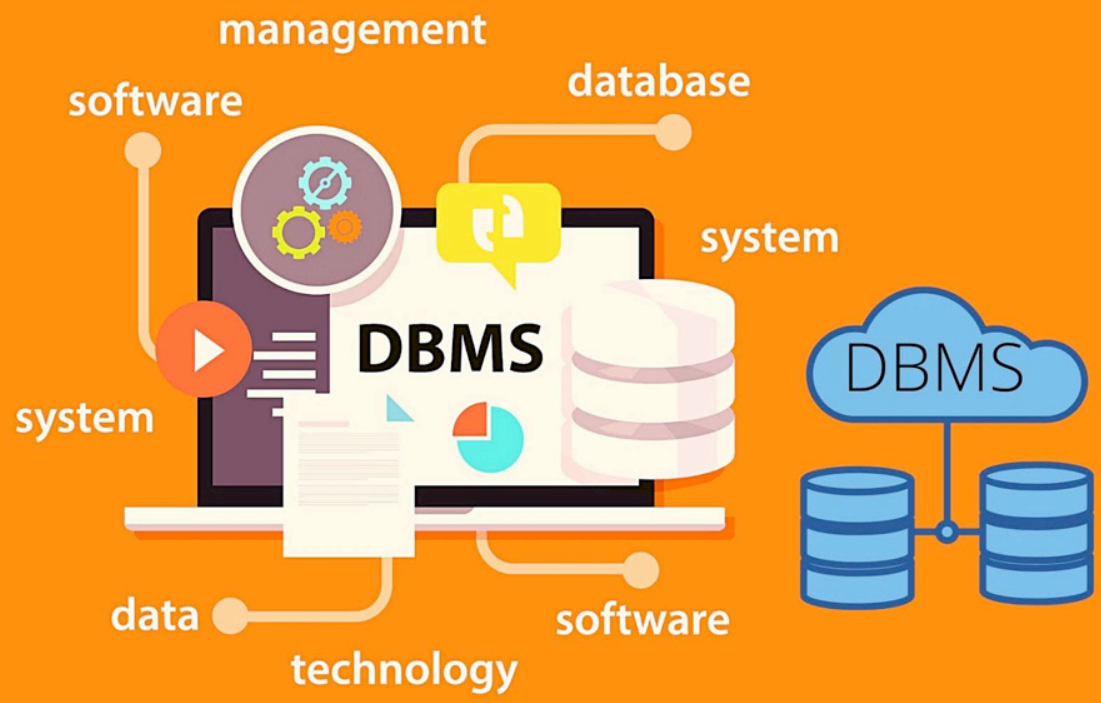
Final projects, homework assignments and lab reports were developed to produce insights in the Data Science program including, but not limited to:

- IST 659: Database Administration
- IST 687: Applied Data Science
- IST 718: Big Data Analytics
- IST 719: Information Visualization

The 7 Learning Objectives of the Applied Data Science Program

The Applied Data Science Program has seven learning objectives which were exemplified by the applications in this portfolio:

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 relevant stakeholder and professionals in their organization.
7. Synthesize the ethical dimensions of data science practice.



IST 659: Database Management

Summer 2018, Final Project: Home Field Advantage
Professor Chad Harper

Link to Class Project

Project Description – IST 659

In the Database Administration course, taught by Chad Harper, I developed a College Football Stadium Database in order to not only organize college teams by their stadium, conference, coordinates, yards per game and points per game, but to also determine if there was some sort of advantage to playing at home, which I did by creating a custom equation of home vs. away stats.

While developing the database, numerous data were collected, and the data were organized and implemented within the database using conceptual (Fig. 1) and logical (Fig. 2) models.

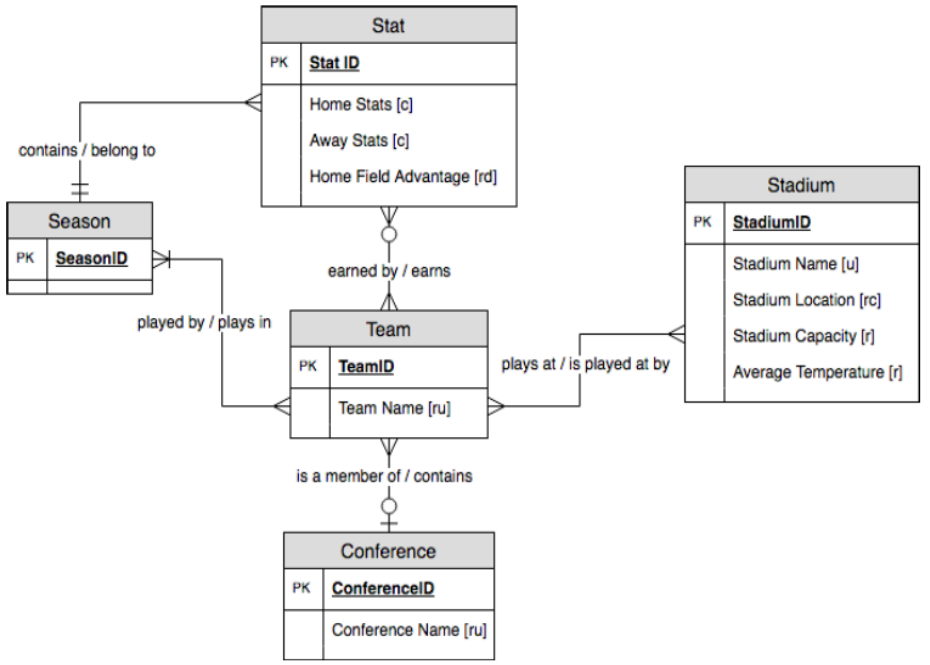


Fig. 1: Conceptual Model

Project Description – IST 659

Microsoft Access was used to populate the data within tables built using SQL Server Management Studio DDL Commands and DML INSERT statements.

I was able to then explore the data and make any necessary changes, using programming objects consisting of views, functions, reports and stored procedures within SQL in order to gain the most insights as efficiently as possible.

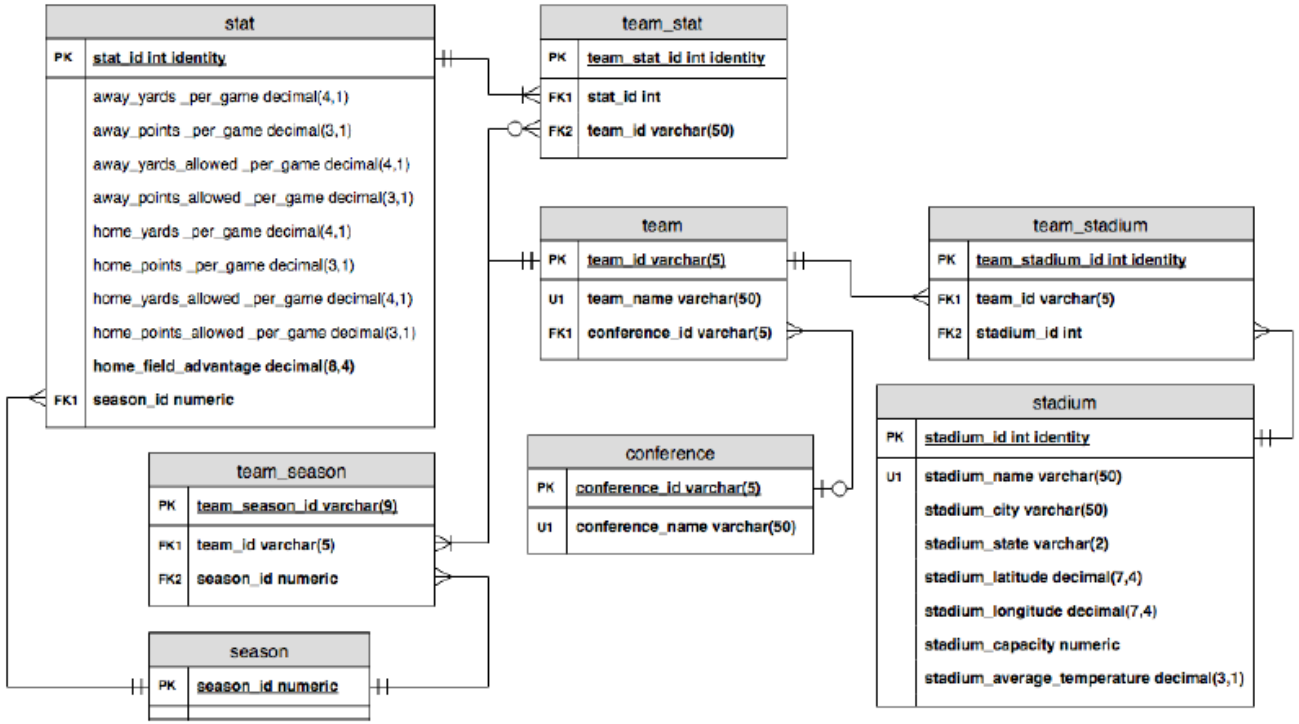


Fig. 2: Logical Model

Project Description – IST 659

My database was able to answer questions such as "which team had the best Home Field Advantage in a single season" and "which team had the best Home Field Advantage over the past 5 years".

These insights were able to help me then expand on my data in order to determine whether or not Home Field Advantage was affected by latitude/longitude, capacity or average temperature (Fig. 3)

team_id	stadium_average_temperature	home_field_advantage
HAWA	86.0	2.24235785
FIU	85.2	0.24996219
FAU	84.8	0.45758234
ARZ	83.8	5.43587582
UTEP	83.8	1.61964289
MIA	83.0	7.42819011
UCF	83.0	7.29387139
USF	82.6	7.97156859
UNLV	82.2	-0.39452108
TEX	80.8	-0.04240357
UCLA	80.6	0.27049601
TXST	80.6	0.68499305
UTSA	80.4	0.55100391
A&M	80.2	4.60249701
FLOR	80.0	7.70571656
FLST	79.8	3.86164945
LALAF	79.8	4.01760456
GASTH	79.4	1.27338436
RICE	79.2	2.19938159
UH	79.2	2.07814984

Fig. 3: Home Field Advantage by Stadium Avg. Temperature

Reflection & Learning Goals – IST 659

This project and course helped me not only design and develop my own data management solution, but also appreciate the nuances and complications that go hand in hand with such implementation vehicles.

Using code in SQL to churn data does not even scratch the surface on what is involved in getting the data in there in the first place, there's also data analysis, database design, data modeling, database management and database implementation.

This course is imperative to the Applied Data Science program but is also a very important stepping-stone to all analysts and data scientists in their relative career paths as they face tougher more complex process such as data warehouses.

Skills Developed – IST 659

Software Learned:

- Microsoft Access
- Microsoft SQL Server

Applied Data Science Program Learning Objective Met:

- “Describe a broad overview of the major practice areas in data science” and “collect and organize data.”

IST 687: Applied Data Science

Summer 2019, Final Project: College Football Blue Bloods

Professor Mohammed Syed

Link to Class Project

Project Description – IST 687

In the Applied Data Science course, taught by Mohammed Syed, I was given weekly assignments which taught data science concepts such as applied statistics, information visualization, text mining and machine learning as well as a final project which involved choosing a real-life dataset in order to use data collection, processing, transformation, management, and analysis to answer questions stemming from that dataset.

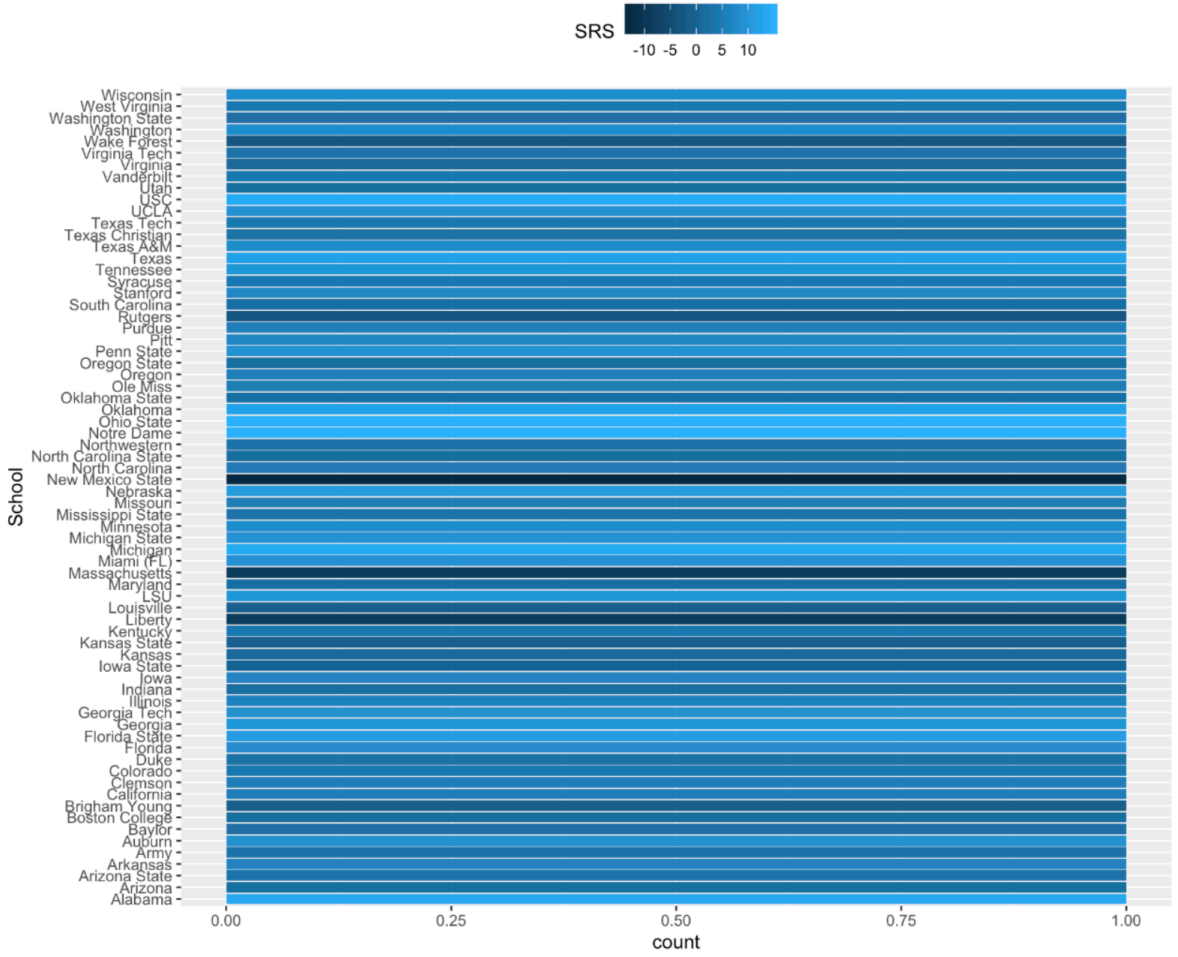


Fig. 4: Simple Rating System (Strength of Schedule/Point Differential rankings) for all teams

Project Description – IST 687

I formed a hypothesis about which College football teams would consistently fall in the Top 10 in the following categories: SRS, SOS, AP, Win%, Conf Win% and Point Differential All-Time.

I ran analysis in order to confirm the hypothesis using "R", the open source statistical analysis and visualization system.

I ran the following Data Analysis and Modeling Functions: sqldf, Rbind, lm, summary and ggplot in R (Fig. 5) as well as t.test, replicate and range functions to measure which schools led in each category.

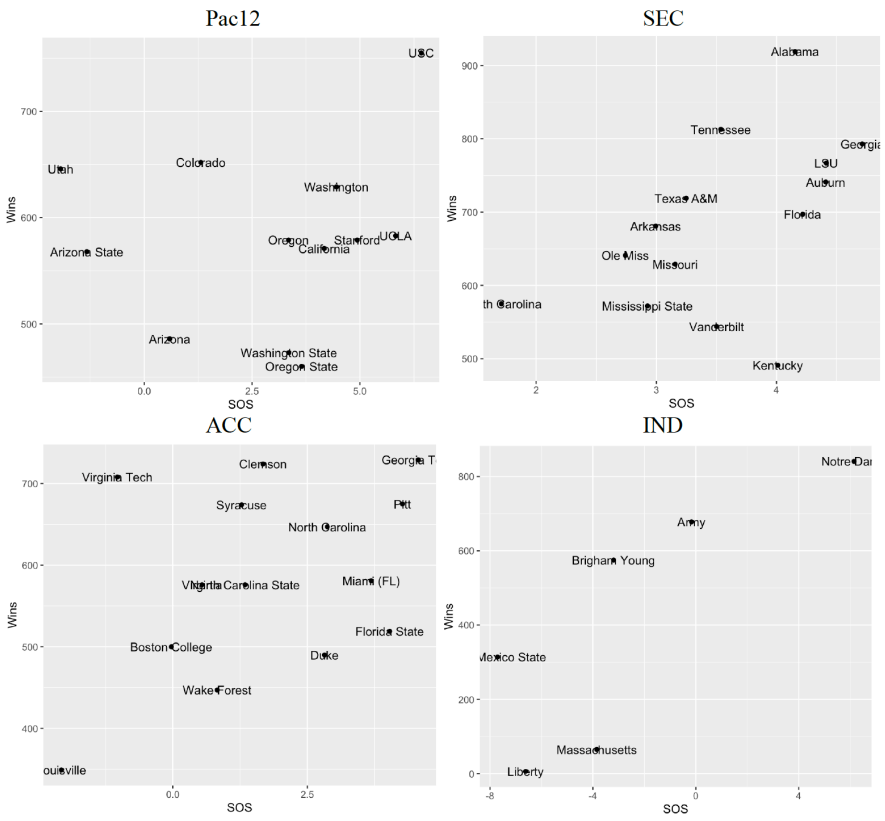


Fig. 5: Strength of Scheduling vs. All-Time Wins

Project Description – IST 687

In the end, my hypothesis was correct and the 8 Blue Blood schools are consistently in the top of every variable (Fig. 6) an all independent variables had a significant correlation with overall wins based on their extremely low p values with SRS being the best individual driver of Wins.

	School	Wins	ConfWins	SOS	SRS	WinPct	ConfPct	APRanks	Diff
25	Ohio State	803	520	5.91392523	14.9603738	0.7378224	0.71578505	107	12.5564486
39	Alabama	919	489	4.15478261	13.4418261	0.7366870	0.61078261	115	13.1179130
66	Notre Dame	841	3	6.14767857	14.7358929	0.7283304	0.01785714	112	11.7824107
27	Michigan	927	512	5.03697674	13.7378295	0.7216279	0.61574419	129	12.7984496
21	Oklahoma	843	502	3.55354545	12.3219091	0.7151273	0.71767273	110	13.7583636
61	USC	755	460	6.42175258	14.1636082	0.6999691	0.71556701	97	10.5157732
15	Texas	857	480	4.07811966	11.8292308	0.6968803	0.61616239	117	10.5431624
32	Nebraska	862	471	3.70831933	10.8342017	0.6885462	0.66474790	119	10.8803361
26	Penn State	865	133	2.19712000	9.1200800	0.6769760	0.13067200	125	9.6900800
47	Tennessee	813	388	3.53701754	10.1019298	0.6745439	0.53785088	114	9.3819298
5	Florida State	519	168	4.03200000	11.3209231	0.6732154	0.32307692	65	10.4386154
49	Georgia	793	382	4.71356522	10.0774783	0.6441391	0.51748696	115	7.3726957
43	LSU	767	345	4.41482456	9.8192105	0.6401754	0.48122807	114	8.0932456
14	Miami (FL)	581	139	3.68719512	8.9825610	0.6315488	0.24150000	82	7.0518293
41	Auburn	741	338	4.41043103	8.7835345	0.6202328	0.43816379	116	6.3022414

Fig. 6: Strength of Scheduling vs. All-Time Wins

Reflection & Learning Goals – IST 687

This course gave way for understanding more essential concepts and characteristics of data such as scripting/code development and data screening, cleaning and linking which gave students experience with R and R-Studio software which is a cornerstone to the data scientist's tool belt.

R is the most popular choice among data analysts worldwide; having knowledge and skill with using it is considered a valuable and marketable job skill for most data scientists.

Organizing and managing data at various stages of a project life-cycle and determining appropriate techniques for analyzing that data are imperative at every level of any profession and this course did an amazing job at preparing me for the future of communicating my findings to decision makers.

Skills Developed – IST 687

Software Learned:

- R
- R-Studio

Applied Data Science Program Learning Objective Met:

- “Identify patterns in data via visualization, statistical analysis, and data mining.”

IST 718: Big Data Analytics

Summer 2020, Lab 1: FBS College Coaches

Professor Jon Fox

Link to Class Project

Project Description – IST 718

In the Big Data Analytics course, taught by Jon Fox, students are given an introduction to analytical processing tools and techniques for information professionals.

Through a set of three vary data and writing intensive lab reports, students are taught to develop a portfolio of resources, demonstrations, recipes, and examples of various analytical techniques.

In one of those labs, I was tasked with determining the starting salary of Syracuse's next head football coach.

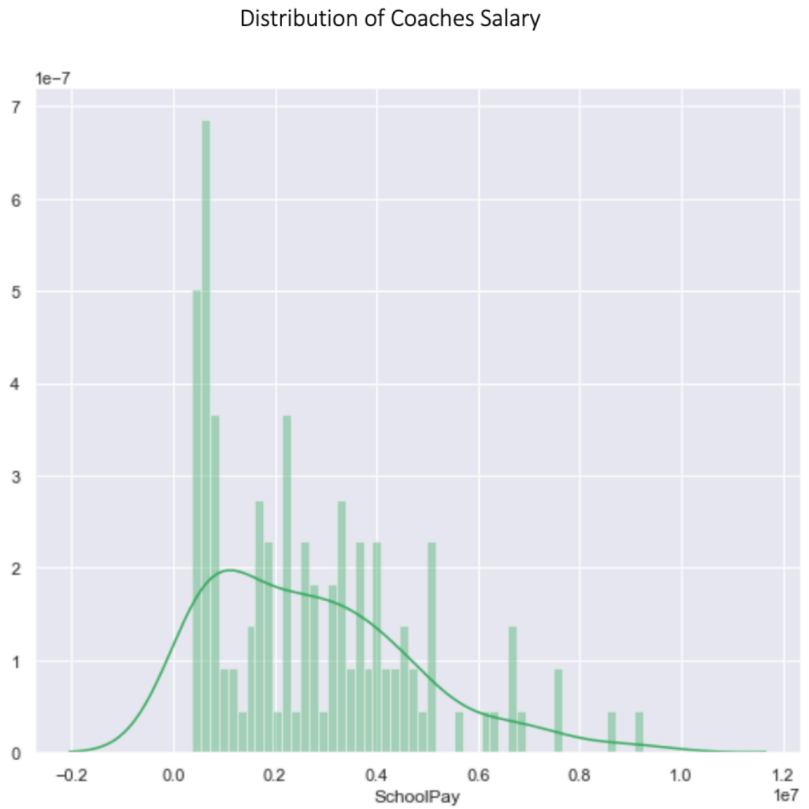


Fig. 7: Distribution of Football Coach Salaries

Project Description – IST 718

I was supposed to complete this task by using a blend of command-line interfaces, quantitative skills and statistics and programming with languages such as R or Python.

In order to learn more software, I chose Python which, like R, is a very popular software among data scientists and analysts.

I had to translate this business challenge into an analytics challenge in order to determine how to attack this unique problem (Fig. 8), which I did by removing fields that lacked a strong correlation with salary.

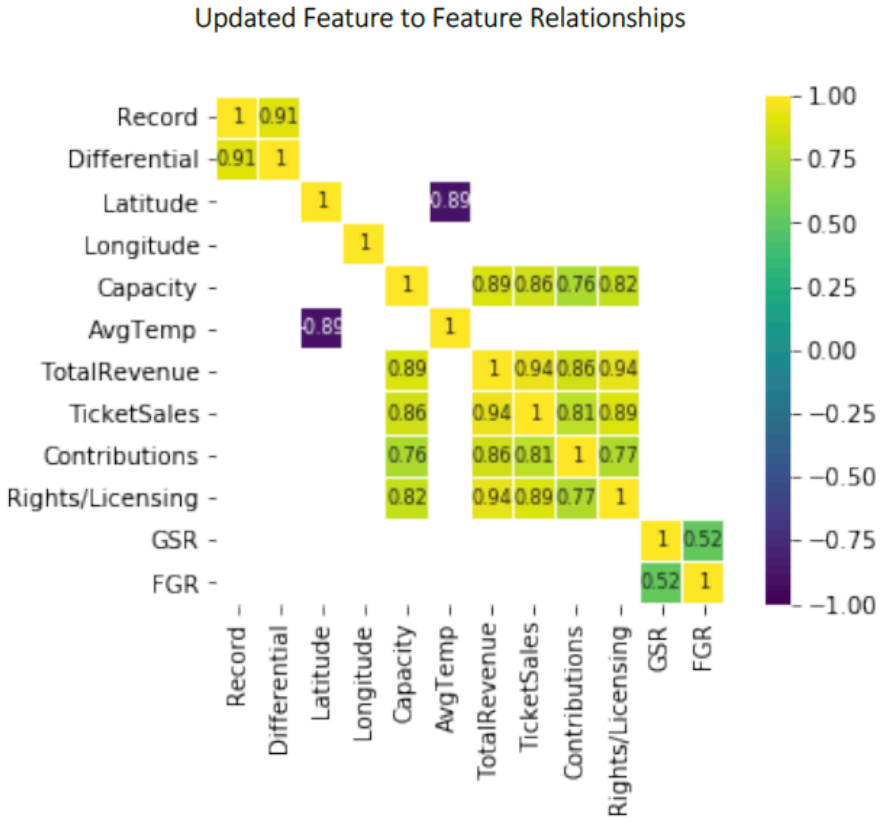


Fig. 8: Feature to feature correlation heatmap

Project Description – IST 718

I used linear and logistic regression to narrow down the most significant fields in order to gain actionable insights (Fig. 9) and (Fig. 10), which in this case was determining the salary of the hypothetical new head football coach at Syracuse.

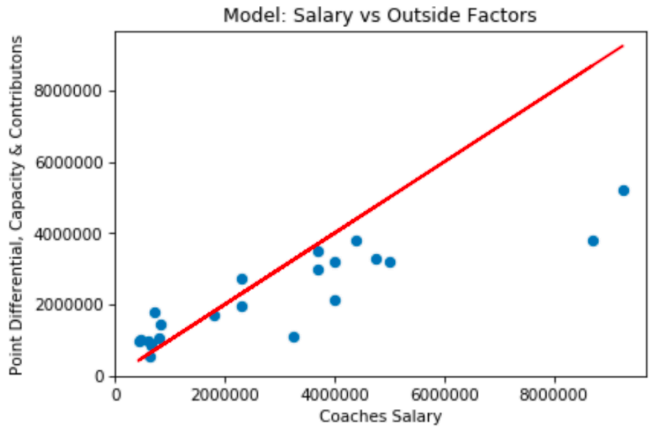


Fig. 9: Plot of my Salary predictions

Second regression model

Dep. Variable:	SchoolPay	**R-squared:**	0.748
Model:	OLS	**Adj. R-squared:**	0.742
Method:	Least Squares	**F-statistic:**	123.0
Date:	Wed, 22 Jul 2020	**Prob (F-statistic):**	1.52e-25
Time:	20:49:27	**Log-Likelihood:**	-1307.7
No. Observations:	86	**AIC:**	2621.
Df Residuals:	83	**BIC:**	2629.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.184e+05	2.76e+05	-0.790	0.432	-7.68e+05	3.31e+05
Capacity	35.3756	7.138	4.956	0.000	21.178	49.573
Contributions	0.0575	0.011	5.186	0.000	0.035	0.080

Omnibus:	16.865	**Durbin-Watson:**	2.423
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	20.573
Skew:	0.971	**Prob(JB):**	3.41e-05
Kurtosis:	4.402	**Cond. No.**	5.98e+07

Fig. 10: Second regression model

Reflection & Learning Goals – IST 718

This course taught me how to pro-actively research solution options vs. relying solely on course textbook content.

Throughout the labs and especially during the final group project, in order to properly scrub data using scripting methods, debug and manipulate data using strictly Python code, it had to be heavily researched which not only instills good habits in a data scientist, but also better prepares an analyst for the real world where problems are not easily solved.

Having to submit labs in report form also improved writing skills as I had to interpret the data, model, analysis, and findings not only in a meaningful way, but also in a way that shareholders, in this case Professor Fox could understand.

Skills Developed – IST 718

Software Learned:

- Python
- Anaconda
- Spark
- Tensorflow
- IBM Watson Studio

Applied Data Science Program Learning Objective Met:

- “Develop alternative strategies based on the data.”
- “Develop a plan of action to implement the business decisions derived from the analyses.”

IST 719: Information Visualization

Spring 2020, Final Project: Which State is the Football State?

Professor Gary Krudys

Link to Class Project

Project Description – IST 719

In the Information Visualization course, taught by Gary Krudys, I developed a portfolio of resources, demonstrations, recipes, and examples of various data visualization skills and techniques through weekly labs and a Final project all while learning and actively using R programming language and Adobe illustrator.

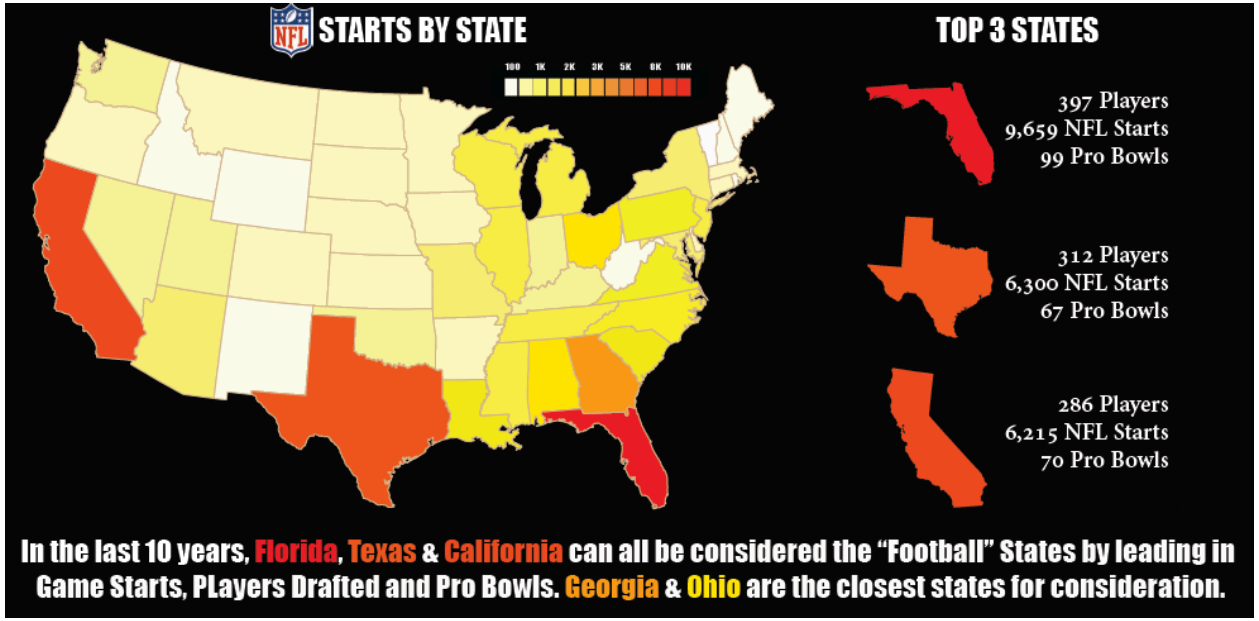


Fig. 11: NFL Starts by State

For the Final Project, I used multiple data cleaning techniques to take control of R graphics' outputs in Illustrator (Fig. 11).

Project Description – IST 719

This resulted in developing custom and unique data plots, visual data exploration charts and design concepts that helped me visually communicate the story in the data to the end user in a way not previously possible through R.

During the course and throughout the labs and project, we as a class and in small groups discussed the various issues related to the ethics of data visualization and were graded using the same guidelines (Fig. 12).

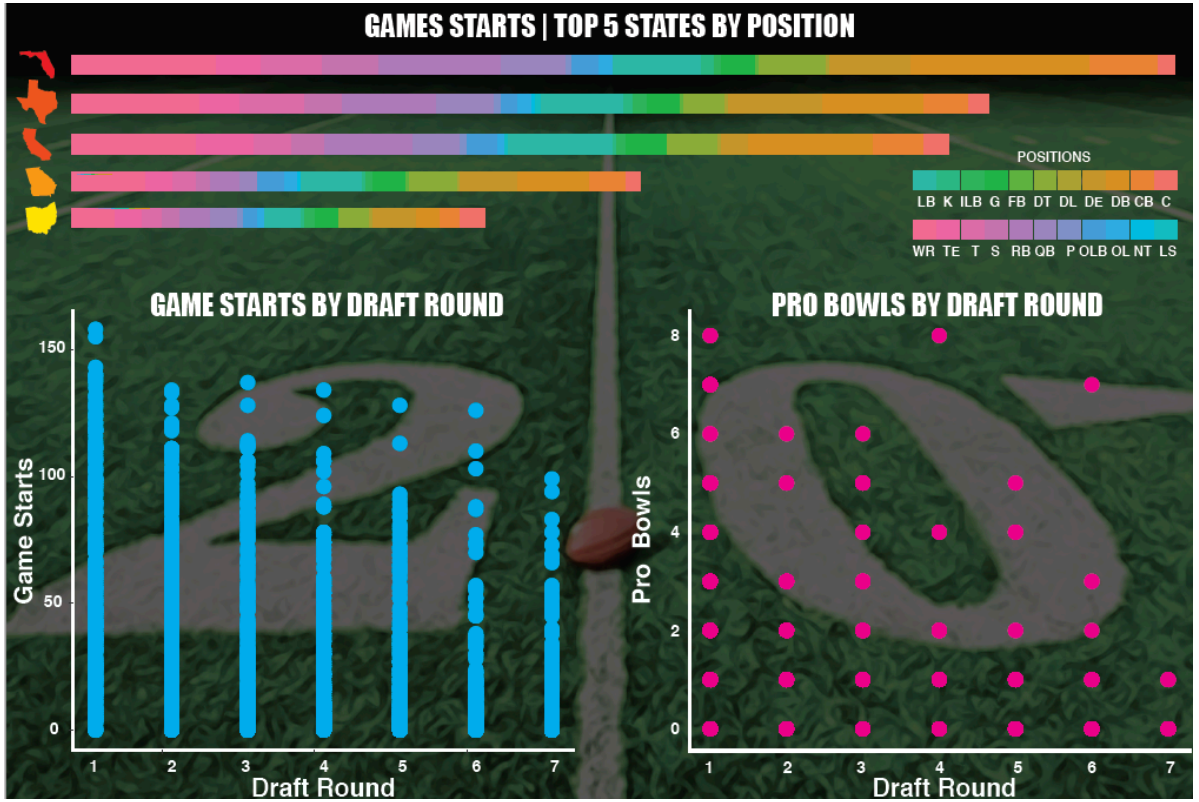


Fig. 12: NFL Starts by State

Reflection & Learning Goals – IST 719

This course taught me how to do basic data cleaning and preparation while using functions inside R to summarize and compare fields, find missing values and subset data. I learned how to use more advanced R techniques which I used throughout the course to create rough plots and identify distributions and relationships in data sets as well as sub-setting and filtering data. Another important lesson learned was how to create rich visual artifacts using Adobe Illustrator to combine R data visualizations, design elements, and context cues while identifying optimal types of visualization to minimize viewer cognitive overload and maximize image interpretability.

On top of the basic set of skills gained from this course, the most important thing I took away was the importance of ethics behind data visualization and how data scientists hold the power to portray the story truthfully and how it is our duty as analysts and scientists in the data field to report accurate and ethically sound data and visualizations.

Skills Developed – IST 719

Software Learned:

- Adobe Illustrator
- R
- Tableau
- Shiny
- Klipfolio

Applied Data Science Program Learning Objective Met:

- "Demonstrate communication skills regarding data and its analysis for relevant stakeholder and professionals in their organization."
- "Synthesize the ethical dimensions of data science practice."

Conclusion

My portfolio has successfully demonstrated the implementation of the learning objectives and the major practice areas that the Data Science program exemplifies:

- fundamentals and key database concepts, database development life cycle, create/design databases and database objects using DBMS and queries in Structured Query Language
- data management and principles and practices in data screening and linking using R and R-Studio as well as processing, aggregation, summarization, searching and the ability to communicate my findings to stakeholders
- obtaining, explaining data structures and data elements, scrubbing data by applying scripting methods for data manipulation in Python and R and exploring/analyzing using qualitative techniques including descriptive statistics, summarization, and visualizations
- data cleaning, controlling the R graphics environment, developing custom plots, visually exploring data, using design concepts to visually communicate the story in the data, and discussing issues related to the ethics of data visualization

References

- Umlang K. E. (n.d.). (2018) IST 659: Database Administration. Retrieved from https://github.com/kyleumlang/MSADS-Portfolio/blob/main/IST%20659%20Database%20Management/Kyle_Umlang_Final_Project_IST659.pdf
- Umlang K. E. (n.d.). (2019) IST 687: Applied Data Science. Retrieved from https://github.com/kyleumlang/MSADS-Portfolio/blob/main/IST%20687%20Applied%20Data%20Science/Kyle_Umlang_Final_Project_IST687.pdf
- Umlang K. E. (n.d.). (2020) IST 718: Big Data Analytics. Retrieved from https://github.com/kyleumlang/MSADS-Portfolio/blob/main/IST%20718%20Big%20Data%20Analytics/Kyle_Umlang_LAB1_IST718.pdf
- Umlang K. E. (n.d.). (2020) IST 719: Information Visualization. Retrieved from https://github.com/kyleumlang/MSADS-Portfolio/blob/main/IST%20719%20Information%20Visualization/Kyle_Umlang_Final_Project_IST719.pdf

Syracuse University

Thank you!

Kyle Umlang
SUID 109166863
keumlang@syr.edu