Syracuse University, School of Information Studies

# Portfolio Milestone

M.S., Applied Data Science

Samantha Brennen-Lisko
SUID: 763650135
https://github.com/sjlisko/MSADS_Portfolio_Milestone

# TABLE OF CONTENTS

**INTROUCTION:**

Syracuse University's School of Information Studies Applied Data Science program gives students the ability to learn how to collect, model, analyze, develop insights, and communicate those findings using data in many different forms. Students learn these skills using different types of techniques and tools, such as R Studio with R code, Spyder and Jupyter notebook with Python, and SQL Server Management Studio with SQL in course such as, Scripting for Data Analysis (Brennen-Lisko, "IST 652,", 2020), Data Admin. Concepts & Database Management (Brennen-Lisko, 659,", 2020), Data Analytics (Brennen-Lisko, "IST 707,", 2020), and Big Data Analysis (Brennen-Lisko, "IST 718,", 2020). The skills developed in the Applied Data Science program furnish data scientist in all fields to add value to any organization, but in the auditing and accounting fields, they can help close the expectations gap between businesses and this accounting and auditing departments.

The Learning Objectives of the Applied Science Data program were demonstrated in the projects and assignments in this portfolio:

1. Describe a broad overview of the major practice areas of data science. (Data Mining, Predictions)
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).

## IST 652: SCRIPTING FOR DATA ANALYSIS:

Project Background:

Under the guidance of Dr. Debbie Landowski, tools and skills of scripting were introduced to solve problems of accessing and preparing data for analysis. This was accomplished using a variety of formats and situations, sometimes known as data wrangling. The scripting learned in this class provided the skills that were used to acquire and clean the data (Fig.1), to examine the data (Fig. 2), and then transform it into visualization and analysis. The final report and presentation focused on both structured and semistructured data to get a better understanding of what types of businesses applied for and revived Personal Paycheck Protection (PPP) loans, what was the amount of the loans, and how many jobs were able to be saved with these loans.

| | LoanRange | BusinessName | Address | City | State | Zip | NAICSCode | BusinessType | RaceEthnicity | Gender | Veteran | NonProfit |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | a $5-10 million | ARCTIC SLOPE NATIVE ASSOCIATION, LTD. | 7000 Uula St | BARROW | AK | 99723.0 | 813920.0 | Non-Profit Organization | Unanswered | Unanswered | Unanswered | Y |
| 1 | a $5-10 million | CRUZ CONSTRUCTION INC | 7000 East Palmer Wasilla Hwy | PALMER | AK | 99645.0 | 238190.0 | Subchapter S Corporation | Unanswered | Unanswered | Unanswered | NaN |
| 2 | a $5-10 million | I. C. E. SERVICES, INC | 2606 C Street | ANCHORAGE | AK | 99503.0 | 722310.0 | Corporation | Unanswered | Unanswered | Unanswered | NaN |
| 3 | a $5-10 million | KATMAI HEALTH SERVICES LLC | 11001 O'MALLEY CENTRE DRIVE, SUITE 204 | ANCHORAGE | AK | 99515.0 | 621111.0 | Limited Liability Company(LLC) | Unanswered | Unanswered | Unanswered | NaN |
| 4 | a $5-10 million | MATANUSKA TELEPHONE ASSOCIATION | 1740 S. CHUGACH ST | PALMER | AK | 99645.0 | 517311.0 | Cooperative | Unanswered | Unanswered | Unanswered | NaN |

***Fig. 1:*** *PPP data before cleaning, (Brennen-Lisko, "IST 652,", 2020).*

| | LoanRange | BusinessName | Address | City | State | Zip | NAICSCode | BusinessType | RaceEthnicity | Gender | Veteran | NonProfit | JobsRetaine |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 661218 | 661210 | 661201 | 661203 | 661218 | 661202.000000 | 654435.000000 | 659789 | 661218 | 661218 | 661218 | 42462 | 620712.00000 |
| unique | 5 | 656592 | 628512 | 15791 | 57 | NaN | NaN | 17 | 7 | 3 | 3 | 1 | Nal |
| top | e $150,000-350,000 | NEW APPLICATION | PO BOX | NEW YORK | CA | NaN | NaN | Corporation | Unanswered | Unanswered | Unanswered | Y | Nal |
| freq | 379054 | 53 | 90 | 12851 | 87689 | NaN | NaN | 275482 | 566717 | 470253 | 521413 | 42462 | Nal |
| mean | NaN | NaN | NaN | NaN | NaN | 51309.965653 | 504693.558508 | NaN | NaN | NaN | NaN | NaN | 50.67811 |
| std | NaN | NaN | NaN | NaN | NaN | 30751.832423 | 177885.044771 | NaN | NaN | NaN | NaN | NaN | 71.19764 |
| min | NaN | NaN | NaN | NaN | NaN | 256.000000 | 111110.000000 | NaN | NaN | NaN | NaN | NaN | 0.00000 |
| 25% | NaN | NaN | NaN | NaN | NaN | 23606.000000 | 337110.000000 | NaN | NaN | NaN | NaN | NaN | 16.00000 |
| 50% | NaN | NaN | NaN | NaN | NaN | 48648.000000 | 541110.000000 | NaN | NaN | NaN | NaN | NaN | 28.00000 |

***Fig. 2:*** *Statistical analysis of PPP data, (Brennen-Lisko, "IST 652,", 2020).*

The next step was to examine current tweets about PPP loans to see what people were saying about the loan program. It was found that there were quite a few positive and negative but the negative tweets seem to get retweeted at a faster rate.

Retweets: 108

Tweet text: RT @SenMcSallyAZ: Over 1 million Arizona jobs were saved by the Paycheck Protection Program! Last Saturday, I visited with @PrescottBrewing owners John & Roxane who shared that they were able to pay their hourly workers using their PPP loan.

Retweets: 5381

Tweet Text: RT @TheRickyDavila: I still can't get over the fact that Steven Mnuchin funneled $500B in PPP loan covid relief to entities of his choice including himself, Devin Nunes, Jared Kushner, Kanye West, Moscow Mitch's wife Elaine Chao, but $600 for struggling Americans is too much. Evil is as evil does. (One of the retweets can be found here: https://twitter.com/MarshaDB54.)

Retweets: 584

Tweet text: RT @KlasfeldReports: Gas station secured small business bailout money, then paid for Trump billboards, via CNNPolitics https://t.co/y7ZqB9G… (The CNN story, which can be found here: https://www.cnn.com/2020/08/28/politics/trump-billboards-ppp-loan-invs/index.html)

***Fig. 3:*** *Tweets collected about PPP Loans, (Brennen-Lisko, "IST 652,", 2020).*

Even though there was quite a bit a controversy with the program, this analysis showed that millions of jobs were able to be saved. (Fig. 4)

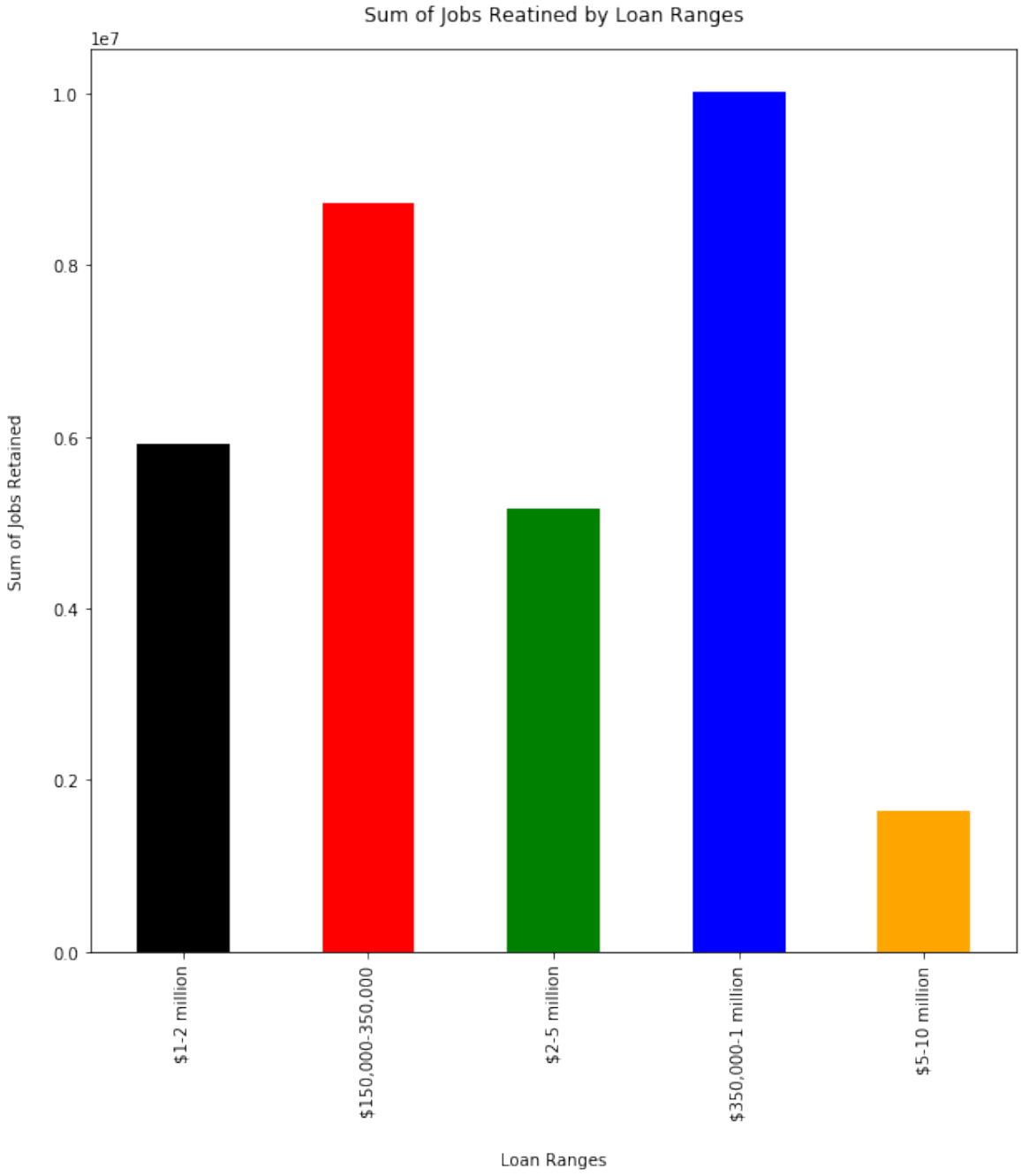



***Fig. 4:*** *Sum of jobs retained with PPP Loans, (Brennen-Lisko, "IST 652,", 2020).*

Skills Achieved:

While this was a great class to begin learning the Python scripting language, one thing I did discover was that it did seem to lack delving into statistical methods of modeling data. While examining the data, there were times I wanted to delve into doing more with linear and logistical regression. While I had learned the skills in R with IST707: Data Analytics, it would not be until IST718: Big Data Analysis, that I began to learn these statistical abilities in Python.

This class and project did give a good base knowledge and preparation for IST718: Big Data Analysis. It should be also noted that these base skills can be used not only with Python applications such as Spyder and Jupyter notebook, but also MS Power BI, which is becoming more widely used in the accounting and auditing fields. As business needs become more depended on data scientists, accounts and auditors, will need to develop the skill to be able to communicate with data scientist, because while they may be speaking about the same thing, the words they use to communicate are from two different languages.

**IST 652: DATA ADMIN. CONCEPTS & DATABASE MANAGEMENT:**

Project Background:

Under the guidance of Dale Thompson, tools and skills needed to build a database for the non-profit called PAGE of Wake County, Inc. (PAGE) was designed (Brennen-Lisko, "IST 659,", 2020). Some of the concerns PAGE has is with how they communicate to key stakeholders, such as the students, parents, and schools that they serve. They needed to increase their communication abilities, if they want to be able to increase the funds, they raise through their Super Saturday classes and membership drive. The next challenge was how to increase their fundraising with local businesses.

Upon discussing these challenges, it was discovered that there was not a central database in which PAGE could see parents' email addresses, home addresses, or phone numbers. Nor did they know all the point of contacts for the businesses that had donated funds to them. PAGE decided to start with creating a database that would capture both the data for Super Saturday and the Wake County Spelling Bee to begin with. Based upon their needs and business rules, a Conceptual and Logical model were built to show the relationships between the students, parents, Super Saturday, the Spelling Bee (Fig. 5). The tables and data in those tables were then created in SQL Server Management Studio from the Excel spreadsheets provided by PAGE. These tables in SQL Server Management Studio where then linked to MS Access and MS Power BI for further analysis (Fig. 6 & 7). This created a powerful tool for PAGE to be able to see how they can filter down the data to one particular student or school thus letting them know where they needed to focus their communications. This will give PAGE, not only increase their services to more students and schools, but also increase their fundraising activities.

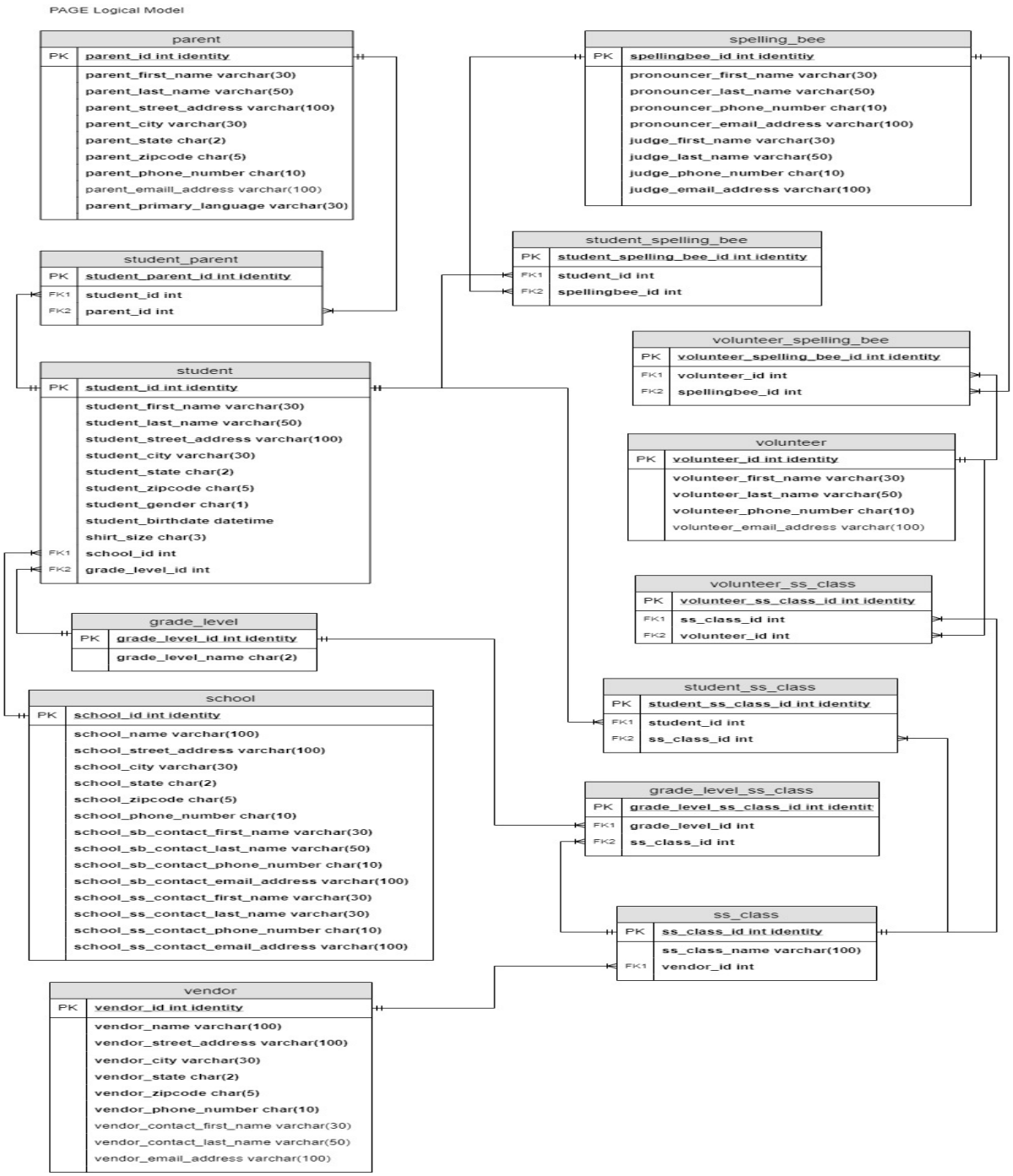
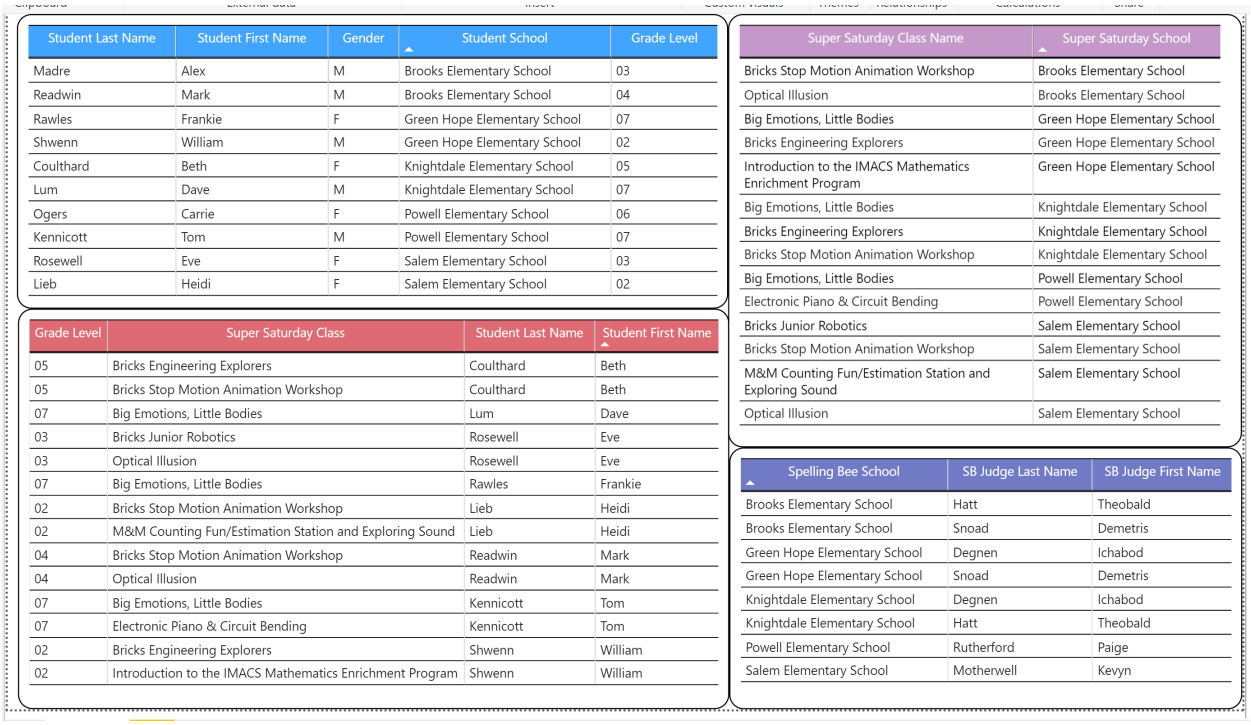



***Fig. 5:*** *PAGE of Wake County Logical Model, (Brennen-Lisko, "IST 659,", 2020).*



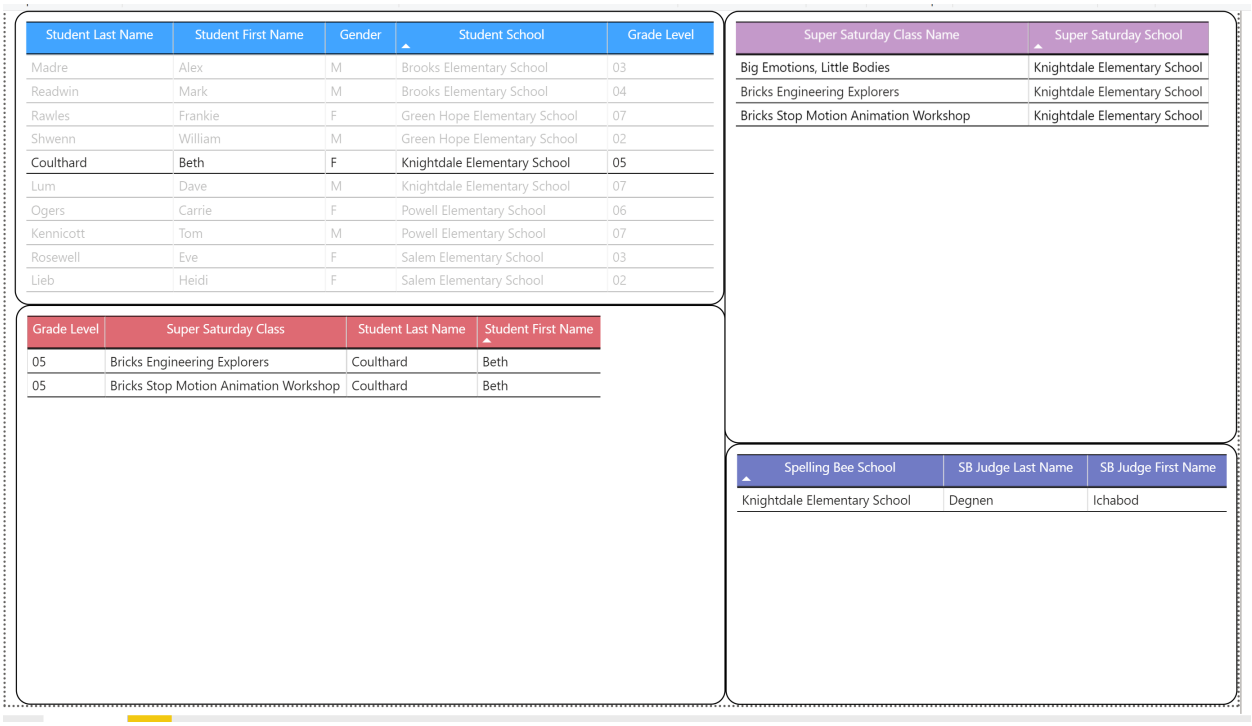
| Student Last Name | Student First Name | Gender | Student School | Grade Level |
|---|---|---|---|---|
| Madre | Alex | M | Brooks Elementary School | 03 |
| Readwin | Mark | M | Brooks Elementary School | 04 |
| Rawles | Frankie | F | Green Hope Elementary School | 07 |
| Shwenn | William | M | Green Hope Elementary School | 02 |
| Coulthard | Beth | F | Knightdale Elementary School | 05 |
| Lum | Dave | M | Knightdale Elementary School | 07 |
| Ogers | Carrie | F | Powell Elementary School | 06 |
| Kennicott | Tom | M | Powell Elementary School | 07 |
| Rosewell | Eve | F | Salem Elementary School | 03 |
| Lieb | Heidi | F | Salem Elementary School | 02 |

| Super Saturday Class Name | Super Saturday School |
|---|---|
| Bricks Stop Motion Animation Workshop | Brooks Elementary School |
| Optical Illusion | Brooks Elementary School |
| Big Emotions, Little Bodies | Green Hope Elementary School |
| Bricks Engineering Explorers | Green Hope Elementary School |
| Introduction to the IMACS Mathematics Enrichment Program | Green Hope Elementary School |
| Big Emotions, Little Bodies | Knightdale Elementary School |
| Bricks Engineering Explorers | Knightdale Elementary School |
| Bricks Stop Motion Animation Workshop | Knightdale Elementary School |
| Big Emotions, Little Bodies | Powell Elementary School |
| Electronic Piano & Circuit Bending | Powell Elementary School |
| Bricks Junior Robotics | Salem Elementary School |
| Bricks Stop Motion Animation Workshop | Salem Elementary School |
| M&M Counting Fun/Estimation Station and Exploring Sound | Salem Elementary School |
| Optical Illusion | Salem Elementary School |

| Grade Level | Super Saturday Class | Student Last Name | Student First Name |
|---|---|---|---|
| 05 | Bricks Engineering Explorers | Coulthard | Beth |
| 05 | Bricks Stop Motion Animation Workshop | Coulthard | Beth |
| 07 | Big Emotions, Little Bodies | Lum | Dave |
| 03 | Bricks Junior Robotics | Rosewell | Eve |
| 03 | Optical Illusion | Rosewell | Eve |
| 07 | Big Emotions, Little Bodies | Rawles | Frankie |
| 02 | Bricks Stop Motion Animation Workshop | Lieb | Heidi |
| 02 | M&M Counting Fun/Estimation Station and Exploring Sound | Lieb | Heidi |
| 04 | Bricks Stop Motion Animation Workshop | Readwin | Mark |
| 04 | Optical Illusion | Readwin | Mark |
| 07 | Big Emotions, Little Bodies | Kennicott | Tom |
| 07 | Electronic Piano & Circuit Bending | Kennicott | Tom |
| 02 | Bricks Engineering Explorers | Shwenn | William |
| 02 | Introduction to the IMACS Mathematics Enrichment Program | Shwenn | William |

| Spelling Bee School | SB Judge Last Name | SB Judge First Name |
|---|---|---|
| Brooks Elementary School | Hatt | Theobald |
| Brooks Elementary School | Snoad | Demetris |
| Green Hope Elementary School | Degnen | Ichabod |
| Green Hope Elementary School | Snoad | Demetris |
| Knightdale Elementary School | Degnen | Ichabod |
| Knightdale Elementary School | Hatt | Theobald |
| Powell Elementary School | Rutherford | Paige |
| Salem Elementary School | Motherwell | Kevyn |

***Fig. 6:*** *PAGE Power BI dashboard, (Brennen-Lisko, "IST 659,", 2020).*



| Student Last Name | Student First Name | Gender | Student School | Grade Level |
|---|---|---|---|---|
| Madre | Alex | M | Brooks Elementary School | 03 |
| Readwin | Mark | M | Brooks Elementary School | 04 |
| Rawles | Frankie | F | Green Hope Elementary School | 07 |
| Shwenn | William | M | Green Hope Elementary School | 02 |
| Coulthard | Beth | F | Knightdale Elementary School | 05 |
| Lum | Dave | M | Knightdale Elementary School | 07 |
| Ogers | Carrie | F | Powell Elementary School | 06 |
| Kennicott | Tom | M | Powell Elementary School | 07 |
| Rosewell | Eve | F | Salem Elementary School | 03 |
| Lieb | Heidi | F | Salem Elementary School | 02 |

| Super Saturday Class Name | Super Saturday School |
|---|---|
| Big Emotions, Little Bodies | Knightdale Elementary School |
| Bricks Engineering Explorers | Knightdale Elementary School |
| Bricks Stop Motion Animation Workshop | Knightdale Elementary School |

| Grade Level | Super Saturday Class | Student Last Name | Student First Name |
|---|---|---|---|
| 05 | Bricks Engineering Explorers | Coulthard | Beth |
| 05 | Bricks Stop Motion Animation Workshop | Coulthard | Beth |

| Spelling Bee School | SB Judge Last Name | SB Judge First Name |
|---|---|---|
| Knightdale Elementary School | Degnen | Ichabod |

***Fig. 7:*** *PAGE Power BI dashboard filtered on the student Beth Coulthard, (Brennen-Lisko, "IST 659,", 2020).*

Skills Achieved:

This project assisted in the developing the skills needed to understand how data is stored and how users access it. This is an imperative skill not only to data scientist who are analyzing the data but also to IT auditors who examine the controls used to mitigate the risk of influencers attacking the integrity of the data. This project also examined the importance of understanding the business rules. For instance, in PAGE's database, the judges and pronouncers were in the Spelling Bee table and all the fields in this table were required. This can create issues as there is only one pronouncer per Spelling Bee but three judges. Had more time been spent communicating with the board, this would have been discovered before the database built, thus averting the challenge of having to make corrections.

This project will help students understand where the data they collect for analyzation is collected, stored, and protected. This will help them understand the importance of ensuring that controls are in place to protect that data, thus giving confidence in their analysis of that data. The skills learned in this class and through this project will not only serve students in the data science industry but also all other industries that work with databases.

**IST 707: DATA ANALYTICS:**

Project Background:

Under the guidance of Dr. Ami Gates, tools and skills of data mining were introduced to solve problems of accessing and preparing data for analysis. The areas of study in which students gained insight include regression, market basket analysis, different classification methods, and clustering. In this group final project, the team decided to look into all areas learned from the class to get a full understanding of what makes the people of one country happier than another (Brennen-Lisko, "IST 707,", 2020). R Studio and R coding was used to conduct analysis using World Happiness Reports (WHR) from the years 2015-2019 from *Kaggle*, a public data mining resource.

Upon examination of the datasets within the WHR, it was discovered that 2015-2017 datasets had 12 or more attributes whereas the 2018-2019 datasets had only nine attributes. The group decided to focus on the nine attributes that were contained on both datasets. Once the datasets had been merged into one dataset, it was spilt into subsets for training (70%) and testing (30%) to be used with the models that require it. A statistical summary was developed to get a better understanding of what was in the data after all the years were combined (Fig. 8).

| | Rank | Score | GDP | Family | Life | Freedom | Corrupt. | Gen. |
|---|---|---|---|---|---|---|---|---|
| **Min** | 1.0 | 2.693 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| **1st Qu** | 40.0 | 4.510 | 0.6065 | 0.8694 | 0.4402 | 0.3098 | 0.0540 | 0.1300 |
| **Median** | 79.0 | 5.322 | 0.9822 | 1.1247 | 0.6473 | 0.4310 | 0.0910 | 0.2020 |
| **Mean** | 78.7 | 5.379 | 0.9160 | 1.0784 | 0.6124 | 0.4111 | 0.1254 | 0.2186 |
| **3rd Qu** | 118.0 | 6.189 | 1.2362 | 1.3273 | 0.8080 | 0.5310 | 0.1560 | 0.2788 |
| **Max** | 158.0 | 7.769 | 2.0960 | 1.6440 | 1.1410 | 0.7240 | 0.5519 | 0.8381 |

***Fig. 8:*** *Statistical Summary of WHR Data, (Brennen-Lisko, "IST 707,", 2020).*

A correlation matrix was also developed to examine if any attributes were highly correlated or if one could overshadow (Fig. 9). This was extremely important to understand before the models were built so that the correct attributes could be used for the models. For this project a Linear Regression model, Support Vector Machines (SVM), Naïve Bayes, and Random Forest models were to see which one was the best predictor of happiness. Of all the models the Linear Regression

model was the best predictor model with an returned an R-squared value of 81.25% and an adjusted R-squared value of 80.90%. This model predicted GDP, family, life, and freedom were the most impactful attributes (Fig. 10).

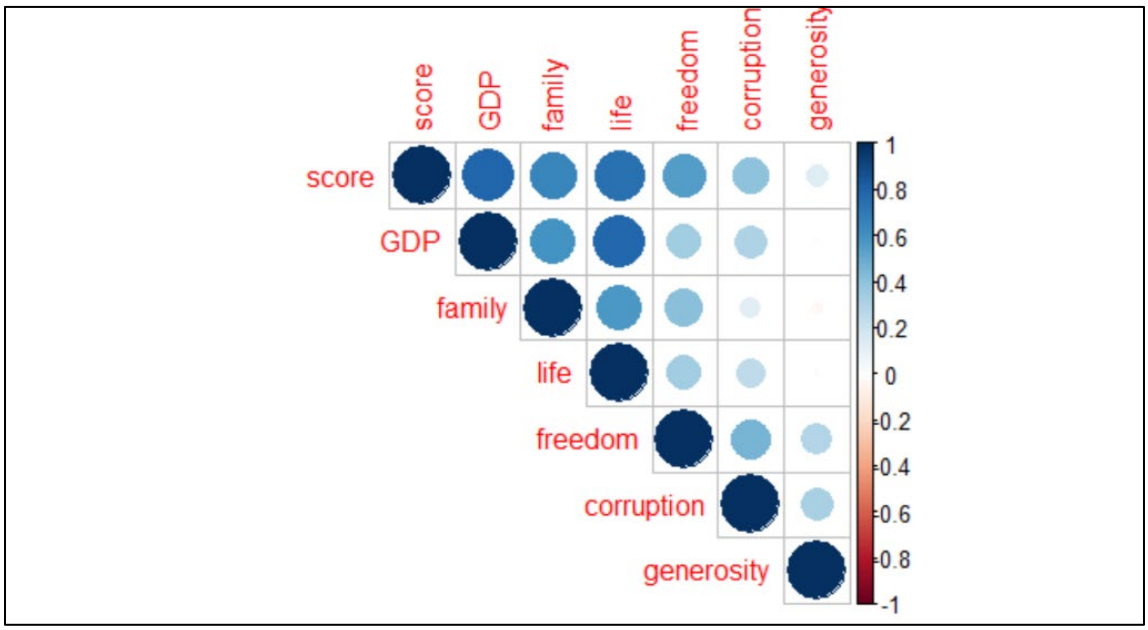



***Fig. 9:*** *Correlation Matrix of Happiness Score Attributes, (Brennen-Lisko, "IST 707,", 2020).*

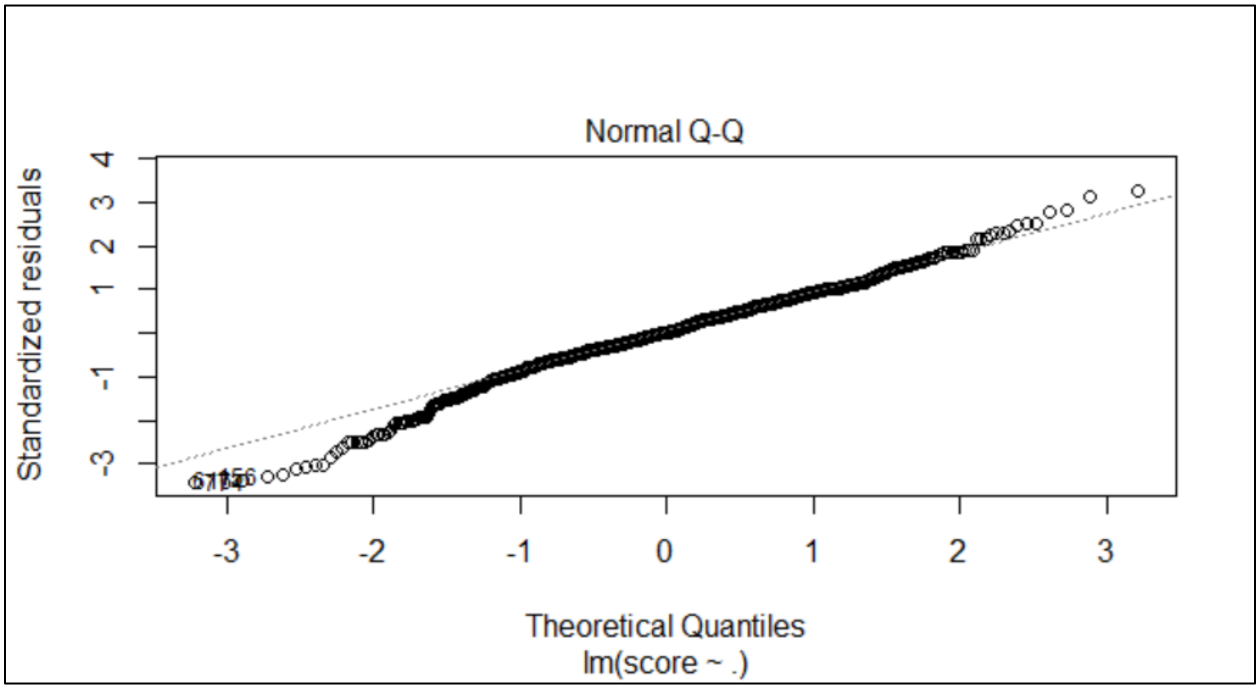



***Fig. 10:*** *Normal Q-Q of WHR Data showing Normal Distribution, (Brennen-Lisko, "IST 707,", 2020).*

Skills Achieved:

This project showed the importance of exploring different models to see which one is the best predictor. For example, the Linear Regression model had R-squared value of 81.25% whereas the Random Forest model had an RMSE of 47.31% when predicting the happiness scores of the test countries. Being able to examine different models and understand the results is an important skill to have for not only the data scientist but also with accountant and auditors. There is an expectations gap between businesses and their accountants and auditors. Data science is used by many businesses to be able to run more efficiently yet the accountants and auditors they work with do not necessarily understand data science nor how to communicate with those data scientists. Skills learned from this project will provide accountants and auditors the needed skills to better understand the direction the business needs to go as well as how to communicate with the data scientists.

**IST 718: BIG DATA ANALYSIS:**

Assignment Background:

Under the guidance of Jon Fox, tools and skills of data mining were introduced to solve problems of accessing and preparing data for analysis. This case study focused on ability to combine datasets and produce meaningful analysis using Python scripting with Jupyter Notebook. This assignment not only looked into combing data but also giving student the ability to decide on what types of data to collect to merge. They were given the data set containing coaches, their university, their salaries, and more (Fig. 11). It was up to the student to decide what they want to add to that dataset, such as stadium size, grade point averages, win/loss records, and more. They then had clean and merge the data sets that did not necessarily match (Fig. 12).

```
(129, 9)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 129 entries, 0 to 128
Data columns (total 9 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   School        129 non-null    object
 1   Conference    129 non-null    object
 2   Coach         129 non-null    object
 3   SchoolPay     129 non-null    object
 4   TotalPay      129 non-null    object
 5   Bonus         129 non-null    object
 6   BonusPaid     129 non-null    object
 7   AssistantPay  129 non-null    object
 8   Buyout        129 non-null    object
```

***Fig. 11:*** *Coaches Dataset, (Brennen-Lisko, "IST 718,", 2020).*

```
(129, 24)
<class 'pandas.core.frame.DataFrame'>
Int64Index: 129 entries, 0 to 90
Data columns (total 24 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   School        129 non-null    object
 1   Conference    129 non-null    object
 2   Coach         129 non-null    object
 3   SchoolPay     129 non-null    int32
 4   TotalPay      129 non-null    int32
 5   Bonus         129 non-null    int32
 6   BonusPaid     129 non-null    int32
 7   AssistantPay  129 non-null    int32
 8   Buyout        129 non-null    int32
 9   Won           129 non-null    int64
 10  Lost          129 non-null    int64
```

```
11  Tied           129 non-null    int64
12  Pct.           129 non-null    float64
13  Years          129 non-null    int64
14  Total Games    129 non-null    int64
15  Stadium        129 non-null    object
16  City           129 non-null    object
17  State          129 non-null    object
18  Capacity       129 non-null    object
19  Cohort Year    129 non-null    int64
20  Year           129 non-null    object
21  Sport          129 non-null    object
22  GSR            129 non-null    int64
23  FGR            126 non-null    float64
```

***Fig. 12:*** *Merged Dataset, (Brennen-Lisko, "IST 718,", 2020).*

With this merged data, the process of creating visualizations to gain insight and understanding could begin. For instance, and Box and Scatterplot was created showing the different salaries of the coaches by the conference their university is in (Fig. 12). This gave insight into which conferences had higher salary averages as well as which conferences have outliers. This showed that conference will be a strong attribute that will need further examination.

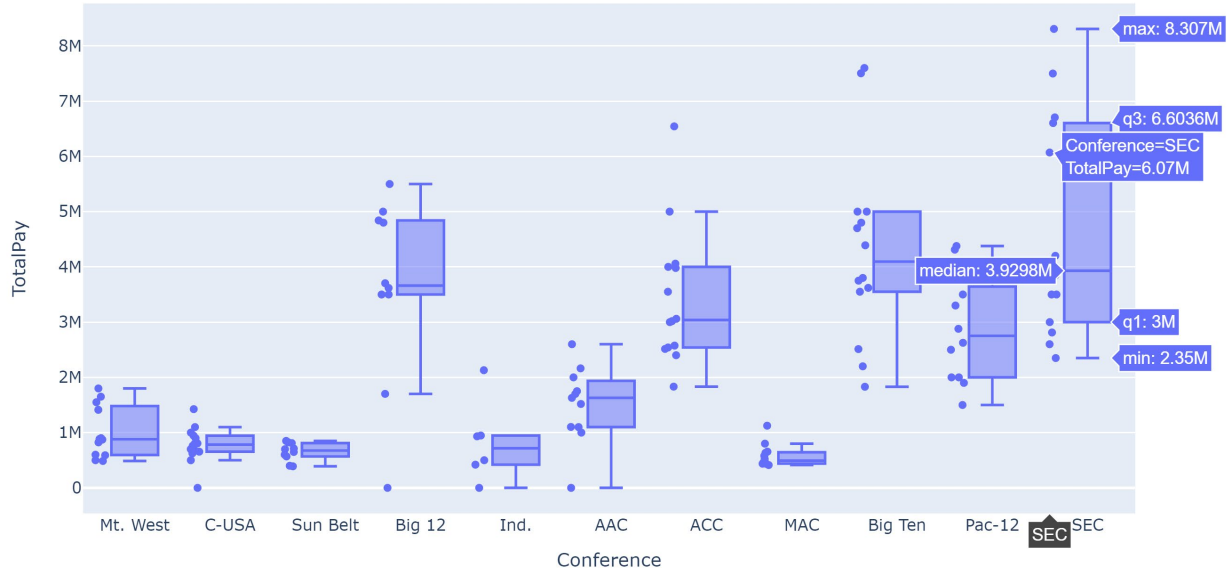



***Fig. 12:*** *Box and Scatterplot of the TotalPay by Conference, (Brennen-Lisko, "IST 718,", 2020).*

Another area that was examined was the wins by total games and conference. This scatterplot also showed that the wins and losses were also a strong attribute that needed to be considered in the model (Fig. 13).

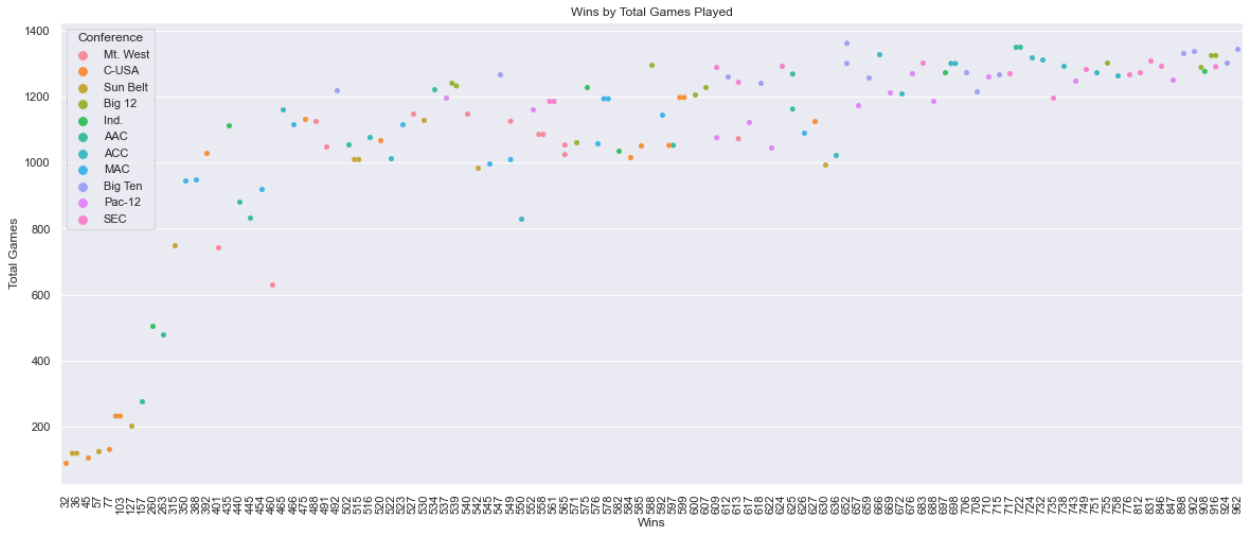



***Fig. 12:*** *Scatterplot of Wins by Total Games and Conference, (Brennen-Lisko, "IST 718,", 2020).*

Once all the attributes had be examined, the building of the Linear Regression model ensued and the conferences and win/loss were the two attributes that gave an R-squared value of 70.1%, which was the best after trying combinations with the grade point average, stadiums, and total games (Fig. 13).

```
Proportion of Test Set Variance Accounted for:  0.719
                            OLS Regression Results
==============================================================================
Dep. Variable:                TotalPay   R-squared:                       0.701
Model:                            OLS   Adj. R-squared:                  0.673
Method:                 Least Squares   F-statistic:                     24.96
Date:                Sat, 17 Oct 2020   Prob (F-statistic):           8.96e-26
Time:                        18:59:00   Log-Likelihood:                -1969.8
No. Observations:                 129   AIC:                             3964.
Df Residuals:                     117   BIC:                             3998.
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
                             coef    std err          t      P>|t|      [0.025      0.
975]
```

```
------------------------------------------------------------------------------
----
Intercept               2.477e+05   4.63e+05      0.535      0.594    -6.7e+05    1.17
e+06
Conference[T.ACC]       1.584e+06   4.47e+05      3.541      0.001    6.98e+05    2.47
e+06
Conference[T.Big 12]    1.664e+06   4.89e+05      3.399      0.001    6.94e+05    2.63
e+06
Conference[T.Big Ten]   2.293e+06   4.58e+05      5.010      0.000    1.39e+06     3.2
e+06
Conference[T.C-USA]    -4.044e+05   4.46e+05     -0.906      0.367   -1.29e+06    4.79
e+05
Conference[T.Ind.]     -8.418e+05   5.54e+05     -1.520      0.131   -1.94e+06    2.55
e+05
Conference[T.MAC]      -9.423e+05   4.54e+05     -2.075      0.040   -1.84e+06   -4.28
e+04
Conference[T.Mt. West] -5.348e+05   4.54e+05     -1.177      0.242   -1.43e+06    3.65
e+05
Conference[T.Pac-12]    1.019e+06   4.64e+05      2.197      0.030       1e+05    1.94
e+06
Conference[T.SEC]       2.635e+06   4.57e+05      5.759      0.000    1.73e+06    3.54
e+06
Conference[T.Sun Belt] -4.086e+05   4.89e+05     -0.835      0.405   -1.38e+06    5.61
e+05
Won                     2457.0509    639.196      3.844      0.000    1191.156    3722
.946
==============================================================================
Omnibus:                        8.693   Durbin-Watson:                   2.103
Prob(Omnibus):                  0.013   Jarque-Bera (JB):               12.142
Skew:                           0.348   Prob(JB):                      0.00231
Kurtosis:                       4.332   Cond. No.                     7.20e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly speci
fied.
[2] The condition number is large, 7.2e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

***Fig. 13:*** *Final Model, (Brennen-Lisko, "IST 718,", 2020).*

Skills Achieved:

This project showed the importance of exploring learning the importance of merging datasets. In today's business world data comes from many different databases within the business and data scientists need to have the skills to be able to merge them and provide meaningful insight and wisdom into that data. Cleaning, merging, and prepping data is one of the most important aspect of the being able to analyze it and this assignment demonstrates those needed skills.

Skills learned from this assignment will provide accountants and auditors the needed ability to better understand how data scientists build that datasets used in analyzing the business. These skills will also enhance the accountants and auditor's ability to communicate with the data scientist thus adding value to any analysis the business needs.

**CONCLUSION:**

The Applied Data Science program gives students the ability to learn how to collect, model, analyze, develop insights, and communicate those findings using data in many different forms and this portfolio is a small demonstration of the skills that students will learn in the program. Learning how data is stored to get a better understanding of the importance of the integrity of that data (Brennen-Lisko, 659,", 2020), gives data scientists the knowledge that what they are communicating will add value to the user of that information. Also having the skills to use not only R coding (Brennen-Lisko, "IST 707,", 2020) but also Python (Brennen-Lisko, "IST 718,", "IST 707,", 2020), are skills that all data scientists need to have to be able to collect, model, analyze, develop insights, and communicate those findings to the users. But in today's market it is not only important for data scientist to have these skills but also other professions in the business world such as people in marketing, supply chain, accounting, and auditing professions.

One area where is a communication challenge is between data scientists and other business profession. For example, when and auditor looks at a spreadsheet, they see rows and columns. When a statistician looks at the same spreadsheet, they see observations and variables. Then when the data scientist looks at it, they see observations and attributes. Now all three people are speaking about the same thing yet they use different words to communicate what they are speaking of. This adds to the expectations gap between auditors, business users, and data scientist to grow even larger, which the opposite direction in which it needs to go. In today's market it is imperative that people become educated in different skillsets and this program provides those needed skillsets.

# REFERENCES


Projects:

Brennen-Lisko, S. (2020). *IST 652: Scripting for Data Analysis.* Retrieved from https://github.com/sjlisko/MSADS_Portfolio_Milestone/tree/main/IST652_Scripting_for_Data_Analysis.

Brennen-Lisko, S. (2020). *IST 659: Data Admin. Concepts & Database Management.* Retrieved from https://github.com/sjlisko/MSADS_Portfolio_Milestone/tree/main/IST659_Data_Admin_Concepts%26Database_Management.

Brennen-Lisko, S. (2020). *IST 707: Data Analytics.* Retrieved from https://github.com/sjlisko/MSADS_Portfolio_Milestone/tree/main/IST707_Data_Analytics.

Brennen-Lisko, S. (2020). *IST 718: Big Data Analytics.* Retrieved from https://github.com/sjlisko/MSADS_Portfolio_Milestone/tree/main/IST718_Big_Data_Analytics.