Syracuse University, School of Information Studies
Master of Science, Applied Data Science

Portfolio Milestone

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

Introduction

Abilities gained:

- Collect
- Model
- Analyze
- Develop Insight
- Communicate Findings

Skills gained from courses:

- IST 652: Scripting for Data Analysis
- IST 659: Data Admin. Concepts & Database Management
- IST 707: Data Analytics
- IST 718: Big Data Analysis

Learning Objectives

- Describe a broad overview of the major practice areas of data science. (Data Mining, Predictions)
- 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 652: Scripting for Data Analysis

- Examine Personal Paycheck Protection (PPP) loans
- Tools and skills of scripting using Python to solve problems
  - Acquire and clean the data
  - Examine the data
  - Transform into visualization and analysis
  - Structured and Semistructured data

Semistructured Data

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.

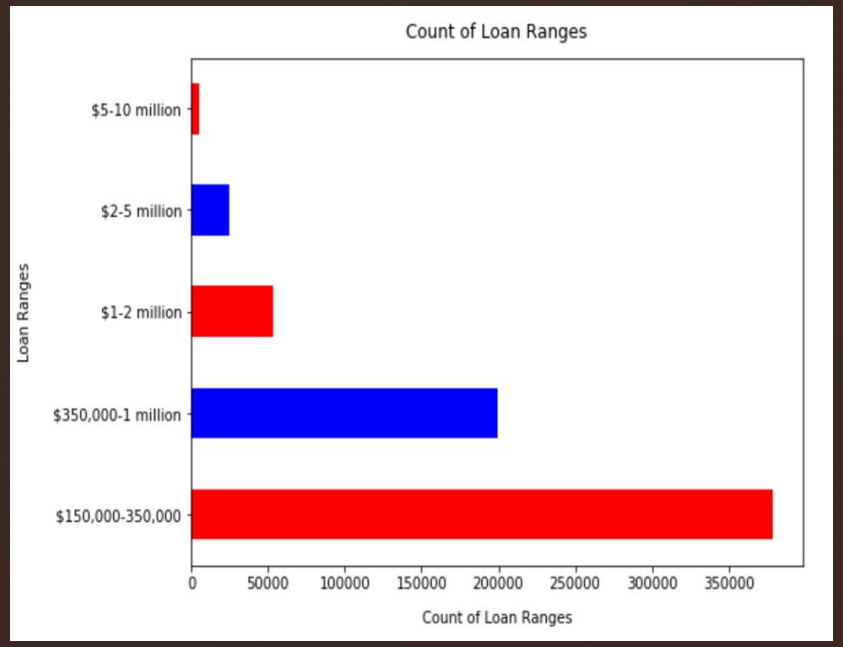
Retweets: 5381

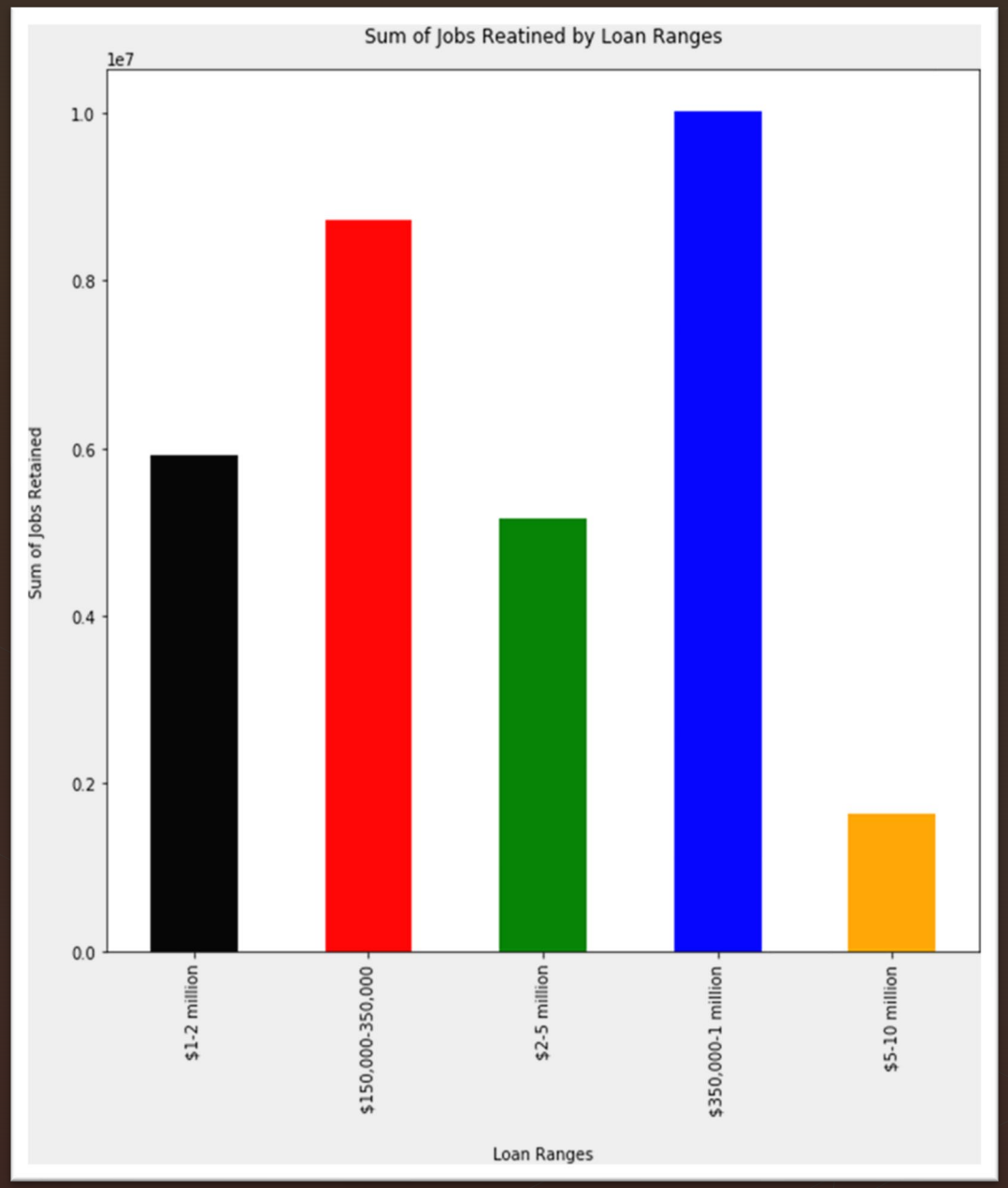
- Tweet Text: RT @TheRickyDavila: I still can't get over the fact that Steven Mnuchin funneled $500B in PPP loan covid relief to entities of his choice including himself, Devin Nunes, Jared Kushner, Kanye West, Moscow Mitch's wife Elaine Chao, but $600 for struggling Americans is too much. Evil is as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

Structured Data Visualizations





Skills Achieved

- Millions of jobs were saved, even with the controversy of the program
- Great class to gain base knowledge of Python
- Skills developed to better communicate
  - Management and Business users
  - Auditors and Accountants

IST 652:Data Administration Concepts & Database Management

- Created database for the non-profit PAGE of Wake County, Inc.
  - Parents
  - Students
  - Schools
  - Volunteers
  - Super Saturday
  - Spelling Bee

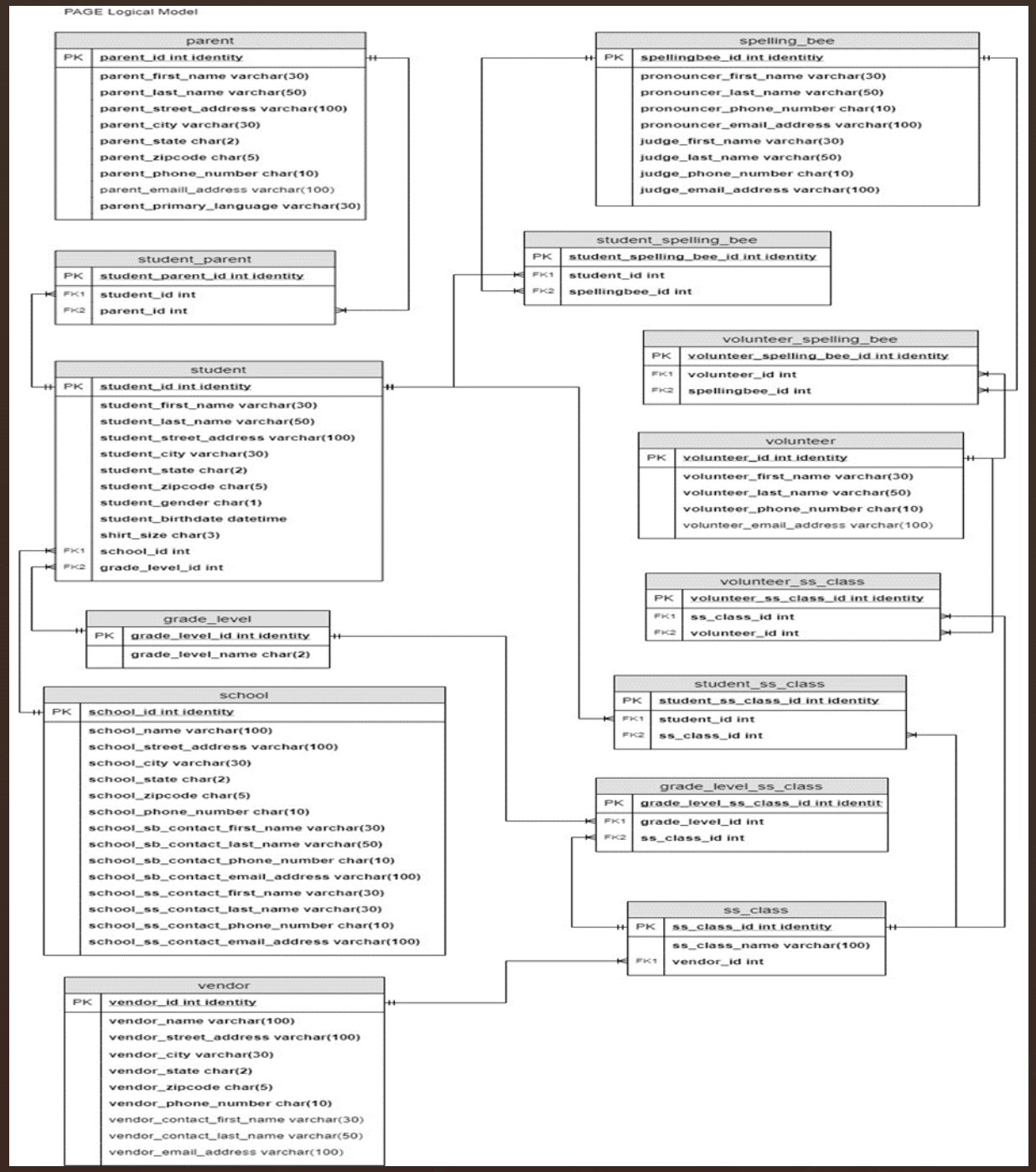
PAGE of Wake County

Table Creation and Reporting

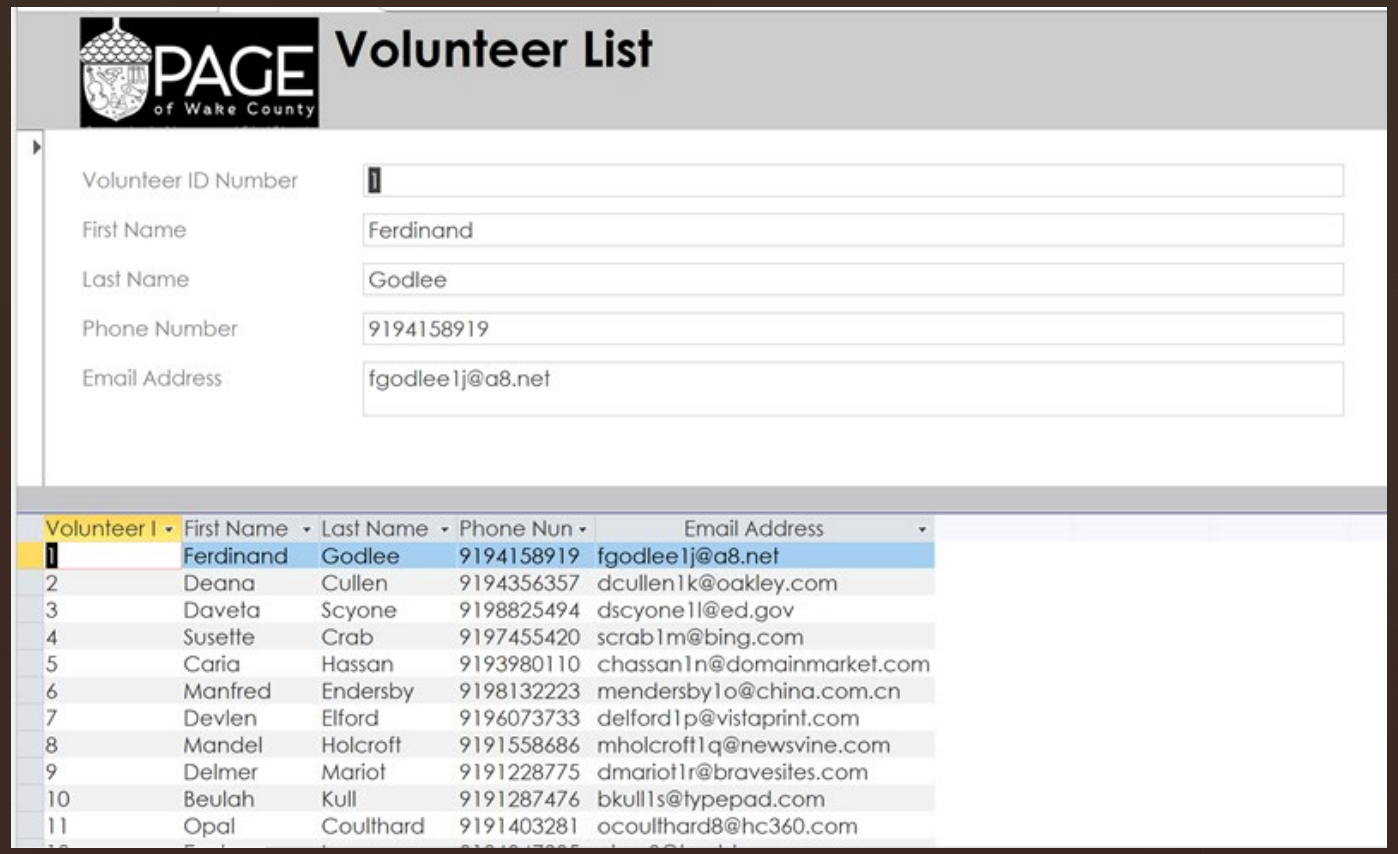
- Conceptual and Logical Models, to design the tables
- SQL Server Management Studio, to create the tables
- Microsoft Access, to create forms and reports
- Microsoft Power BI , to create dashboards

Logical Model and Parent Table from Logical Model

parent	
PK	parent_id int identity
	parent_first_name varchar(30)
	parent_last_name varchar(50)
	parent_street_address varchar(100)
	parent_city varchar(30)
	parent_state char(2)
	parent_zipcode char(5)
	parent_phone_number char(10)
	parent_emaill_address varchar(100)
	parent_primary_language varchar(30)



MS Access Form



Volunteer I	First Name	Last Name	Phone Nun	Email Address
1	Ferdinand	Godlee	9194158919	fgodlee1j@a8.net
2	Deana	Cullen	9194356357	dcullen1k@oakley.com
3	Daveta	Scyone	9198825494	dscyone1l@ed.gov
4	Susette	Crab	9197455420	scrab1m@bing.com
5	Caria	Hassan	9193980110	chassan1n@domainmarket.com
6	Manfred	Endersby	9198132223	mendersby1o@china.com.cn
7	Devlen	Elford	9196073733	delford1p@vistaprint.com
8	Mandel	Holcroft	9191558686	mholcroft1q@newsvine.com
9	Delmer	Mariot	9191228775	dmariot1r@bravesites.com
10	Beulah	Kull	9191287476	bkull1s@typepad.com
11	Opal	Coulthard	9191403281	ocoulthard8@hc360.com

MS Power BI Dashboard

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

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

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

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

Skills Achieved

- Developed skills building a database
  - Storing Data
  - Accessing Data
  - Understanding Data Integrity
  - Develop better communication with Business User
    - Business Rules

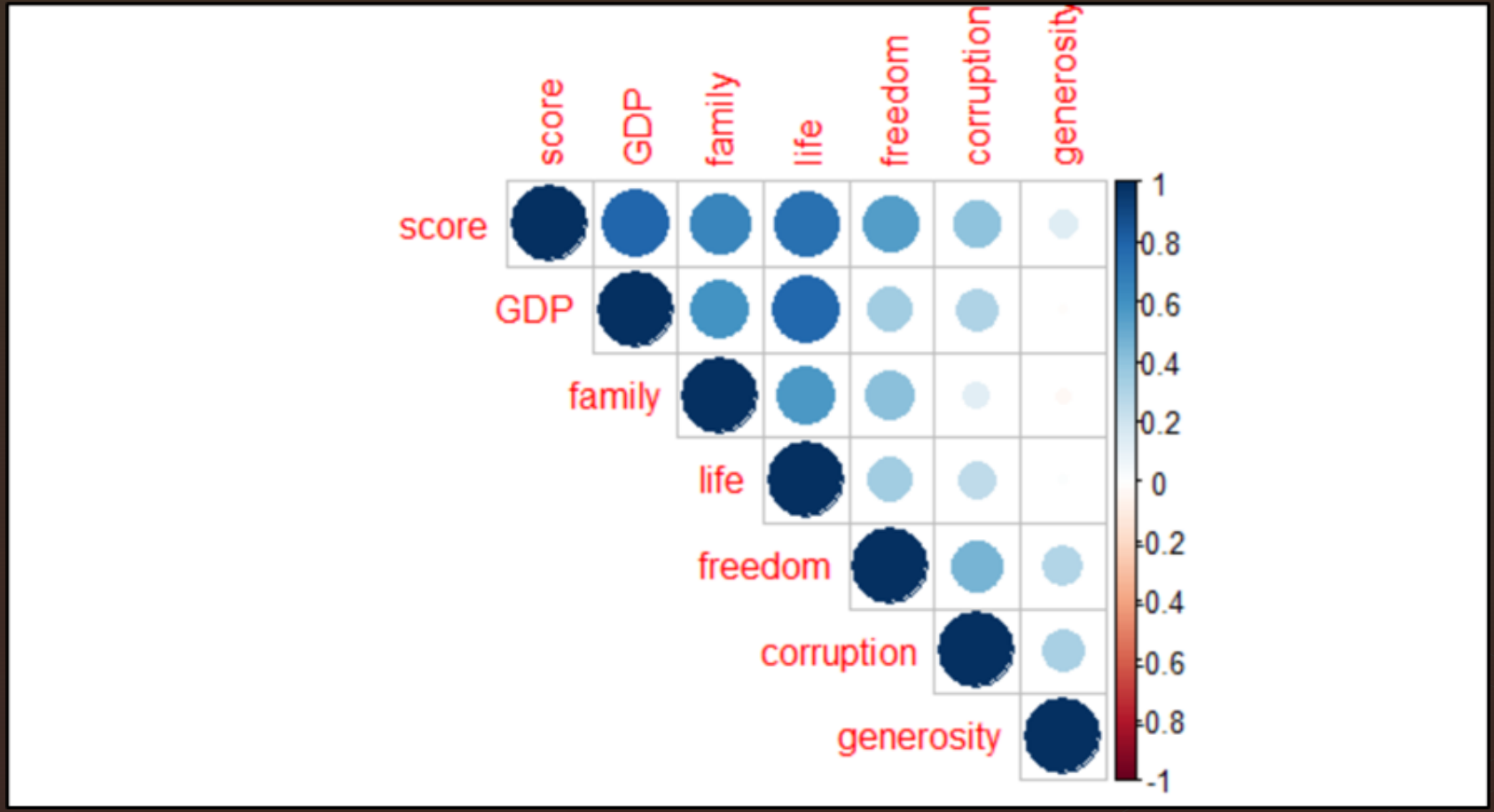
IST 707: Data Analytics

- Examine World Happiness Reports (WHR) from the years 2015-2019
- Tools and skills of data mining using R to solve problems
  - Acquire and clean the data
  - Examine the data
  - Transform into visualization and analysis

Statistical Summary of WHR Data

	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

Correlation Matrix of Happiness Score Attributes



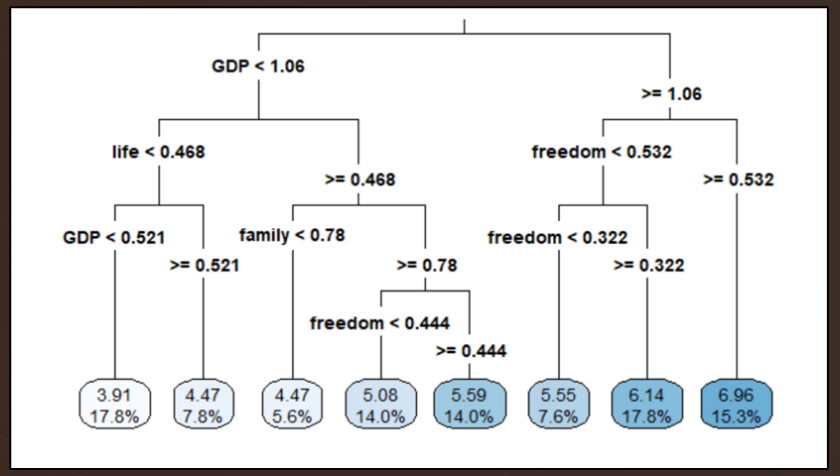
Models

Naïve Bayes Model

Predicted Group	Actual Group: 3 to 4	4 to 5	5 to 6	6 to 7	7 to 8
3 to 4	2	3	0	0	0
4 to 5	0	9	3	0	0
5 to 6	0	1	10	3	0
6 to 7	0	0	1	6	2
7 to 8	0	0	0	1	2

Prediction accuracy rate of 67.44% on the test data set

Random Forest Model



RMSE of 47.31% when predicting the happiness scores of the test countries

Models

Linear Regression Model

Happiness Score = 2.1712 + 1.1439GDP + 0.7010Family + 0.9613Life + 1.2397Freedom + 0.7731Corruption + 0.9049Generosity

R-squared value of 81.25% and an adjusted R-squared value of 80.90%

Support Vector Machines (SVM) Model

		Actual Group				
		3 to 4	4 to 5	5 to 6	6 to 7	7 to 8
Predicted Group	3 to 4	2	0	0	0	0
	4 to 5	3	9	0	0	0
	5 to 6	0	3	14	3	1
	6 to 7	0	0	0	5	1
	7 to 8	0	0	0	1	1

Prediction accuracy rate of 72.09% on the test data set

Skills Achieved

- Importance of exploring different models
  - Linear Regression model had R-squared value of 81.25% whereas the Random Forest model had an RMSE of 47.31%
- Importance of understanding and communicating the models
  - Business Users and Data Scientists
  - Accountants and Auditors

IST 718: Big Data Analysis

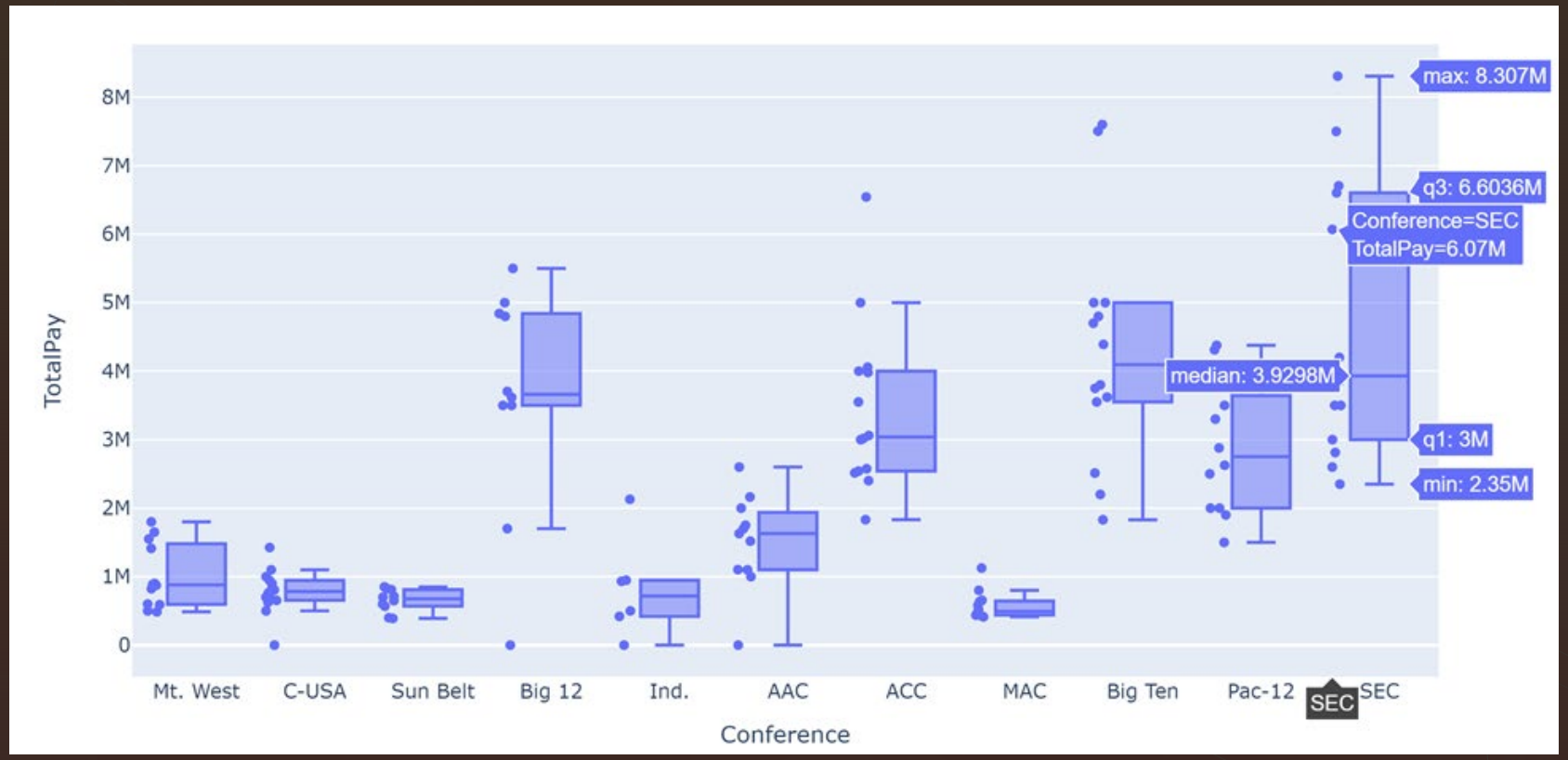
- Examine and add datasets onto Coaches Dataset
- Builds on the tools and skills of scripting using Python from IST 652
  - Acquire and clean the data
  - Combine multiple datasets
  - Examine the data
  - Create models
  - Transform into visualization and analysis
  - Structured data

Datasets

- Decided datasets
  - Coaches
  - Stadium Sizes
  - Grade Point Averages
  - Win/Loss Records

```
(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
```

Coaches Pay by Conference



Final Model

Attributes chosen:

- Conference
- Win Record

```
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.
```

Skills Achieved

- Importance of cleaning, merging, and prepping datasets of difference sizes
- Providing meaningful insight through creating moels in Python
- Increase knowledge thus providing better communication to all users

Conclusion

Abilities gained:

- Collect and merge datasets
- Develop models
- Analyze data
- Develop insight into the data
- Communicate findings in a way all users understand