|  |
| --- |
| Lab Assignment #1  William Jones  Syracuse University  IST 718 |

**Research Question**

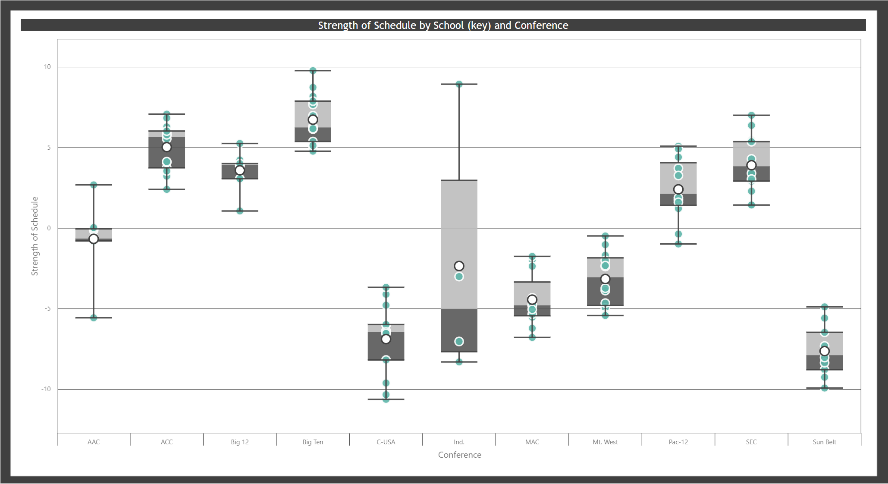
*How can we recommend the best salary (totalpay or schoolpay – your choice) for our next head football coach?*

Specifics:

* *What is the recommended salary for the Syracuse football coach?*
  + Utilizing the model detailed below, I would recommend a TotalPay of ~$3.5-3.6m
* *What would his salary be if we were still in the Big East? What if we went to the Big Ten?*
  + Utilizing the model detailed below, if we were still in the Big East, I would recommend a TotalPay of ~$2.2-2.3m.
  + Utilizing the model detailed below, if we were to move to the Big Ten, I would recommend a TotalPay of ~$4.0-4.1m.
* *What schools did we drop from our data and why?*
  + Schools with $0 TotalPay were removed from the analysis. These would have skewed the analysis.
* *What effect does graduation rate have on the projected salary?*
  + Graduation rate has little to no correlation with Wins or Projected Salary and therefore removed from the recommended salary calculation.
* *How good is our model?*
  + The model has an R2 around ~.750. While this isn’t ideal, it is definitely strong enough to make prediction.
* *What is the single biggest impact on salary size?*
  + The single biggest impact on salary is conference. The Power 5 conferences get a significantly higher salary than non-Power 5 conferences.
* *Develop a geographic visualization that in your view best depicts the conferences’ median salary*
  + See detailed analysis of visualizations created below. Box & Whisker plot created specifically for this requirement.
* *Fit a hierarchical model based on conference*
  + The model includes conference data as part of the prediction. I reviewed multiple different modeling options within Orange Data Mining and the linear regression had the best results.

**Assumptions**

* The 2017 season has just been completed and we (Syracuse University) have just fired our football coach.
* 2018 salaries are “based” upon 2017 season results.
* We are going to target a winning coach. As there are typically 12-14 games in a college season our target coach would have 8 wins & 5 losses in the model. Additionally, we want a coach that pushes our program to play high profiled opponents and therefore we will use a Strength of Schedule at the 3rd Quartile mark for that respective conference. ACC = 6.03, Big East (AAC) = -0.04, & Big Ten = 7.89



**Objective 1** – **O**btain

*Obtain data and understand data structures and data elements.*

“Coaches.csv”

* Source: GitHub assignment repository
* Understand the data elements:
  + Tabular file
  + Removed all columns except for:
    - School: Text (primary key)
    - Conference: Text
    - Coach: Text
    - TotalPay: Numeric

“Graduation Success Rate.csv”

* Source: <https://web3.ncaa.org/apsearch/gsrsearch>
* Understand the data elements:
  + Tabular file
  + Removed all columns except for:
    - School: Text (primary key)
    - GSR: Numeric
    - FGR: Numeric

“College Football Stadium Size.csv”

* Source: <https://www.collegegridirons.com/comparisons-by-capacity/>
* Understand the data elements:
  + Tabular file
  + Removed all columns except for:
    - School: Text (primary key)
    - Stadium: Text
    - Capacity: Numeric

“2017 Team Records.csv”

* Source: <https://www.sports-reference.com/cfb/years/2017-standings.html>
* Understand the data elements:
  + Tabular file
  + Removed all columns except for:
    - School: Text (primary key)
    - W: Numeric
    - L: Numeric
    - Strength of Schedule: Numeric
    - AP High: Numeric
    - Ranked: text – calculated column to show if the school was “Ranked” vs “Not Ranked” during the year.

**Objective 2** – **S**crub

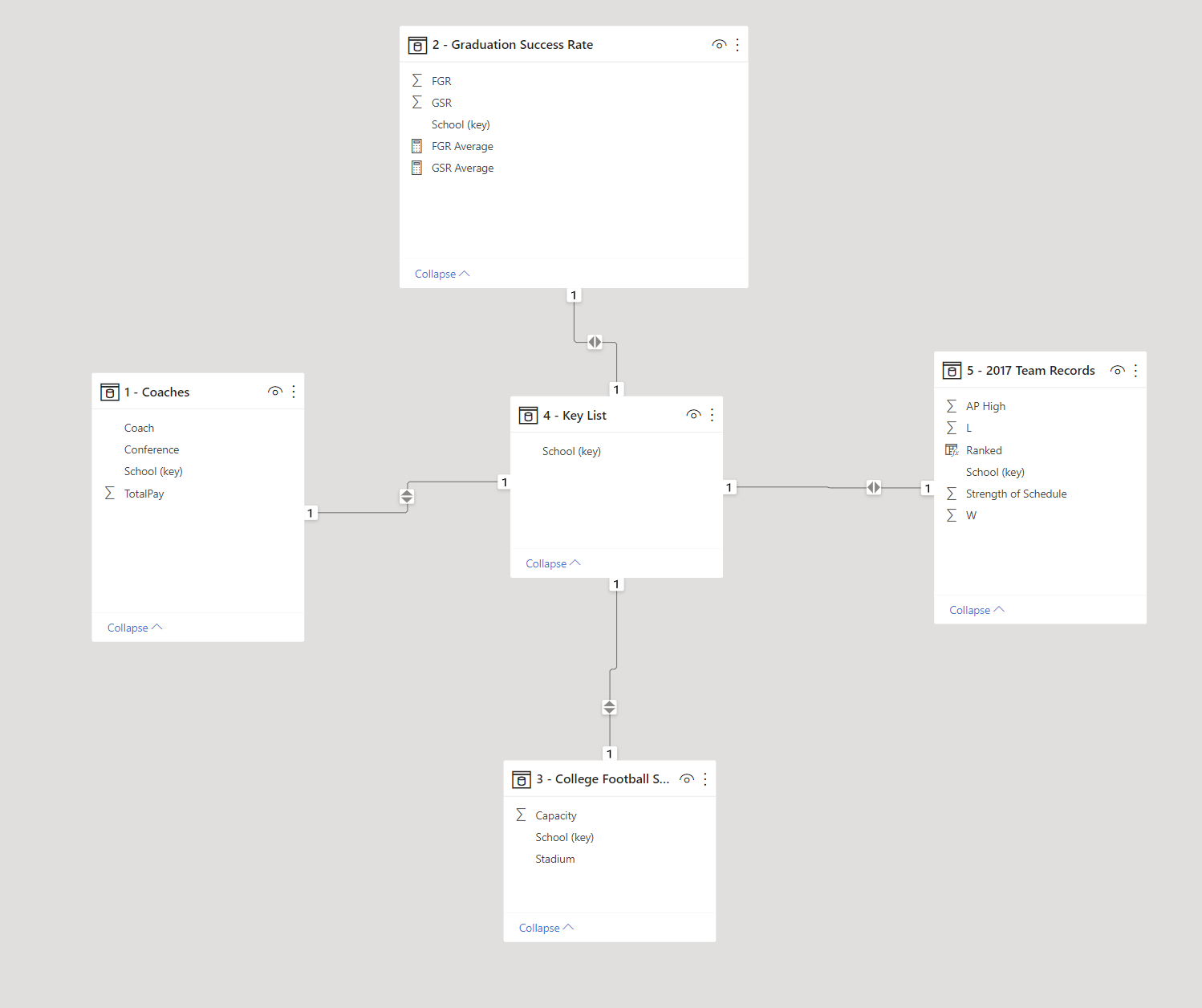
*Scrub data using scripting methods, to include debugging, for data manipulation in R and other tools. Build a data frame for your analysis*

To complete the data cleaning aspect of Lab 1 I utilized PowerBI. PowerBI utilizes PowerQuery to transform data prior to be loading into the dashboard. Many of the transformations are done via no-code functionality but with each action a line of MQuery language is created. The PowerBI and txt formats of the query transformations are attached.

Specific cleaning of note:

* The Coaches.csv data had “--” where there was no pay information. This was replaced with $0. Filtered out at the data visualization level and removed from the testing and training data. Additionally, the pay information came into the engine in multiple formats (ie. “885000” vs “$411,000”). When the data type was selected as Decimal Number in PowerBI these inconsistencies were handled. Liberty was removed as FCS in 2017
* Stadium Sizes and Records data had school names that were inconsistent with the Coaches.csv. Utilizing the Coaches.csv as the “central source of truth”, the other file school names were updated to match. Filtered out Gillette Stadium for UMass as this was not their true home stadium (Mcguirk).
* Graduation Success Rate.csv had school names that were inconsistent with the Coaches.csv. Utilizing the Coaches.csv as the “central source of truth”, the other file school names were updated to match. Temple, Presbyterian, and Tarleton State were removed as not on the Coaches.csv
* All data had the primary key field trimmed and capitalized.

The PowerBI data model looked like this.



Training & Testing Data

All data was included in the modeling phased. To test the model, 10-fold cross validation and 1/3 random sampling were used. See commentary on results in Objective 5.

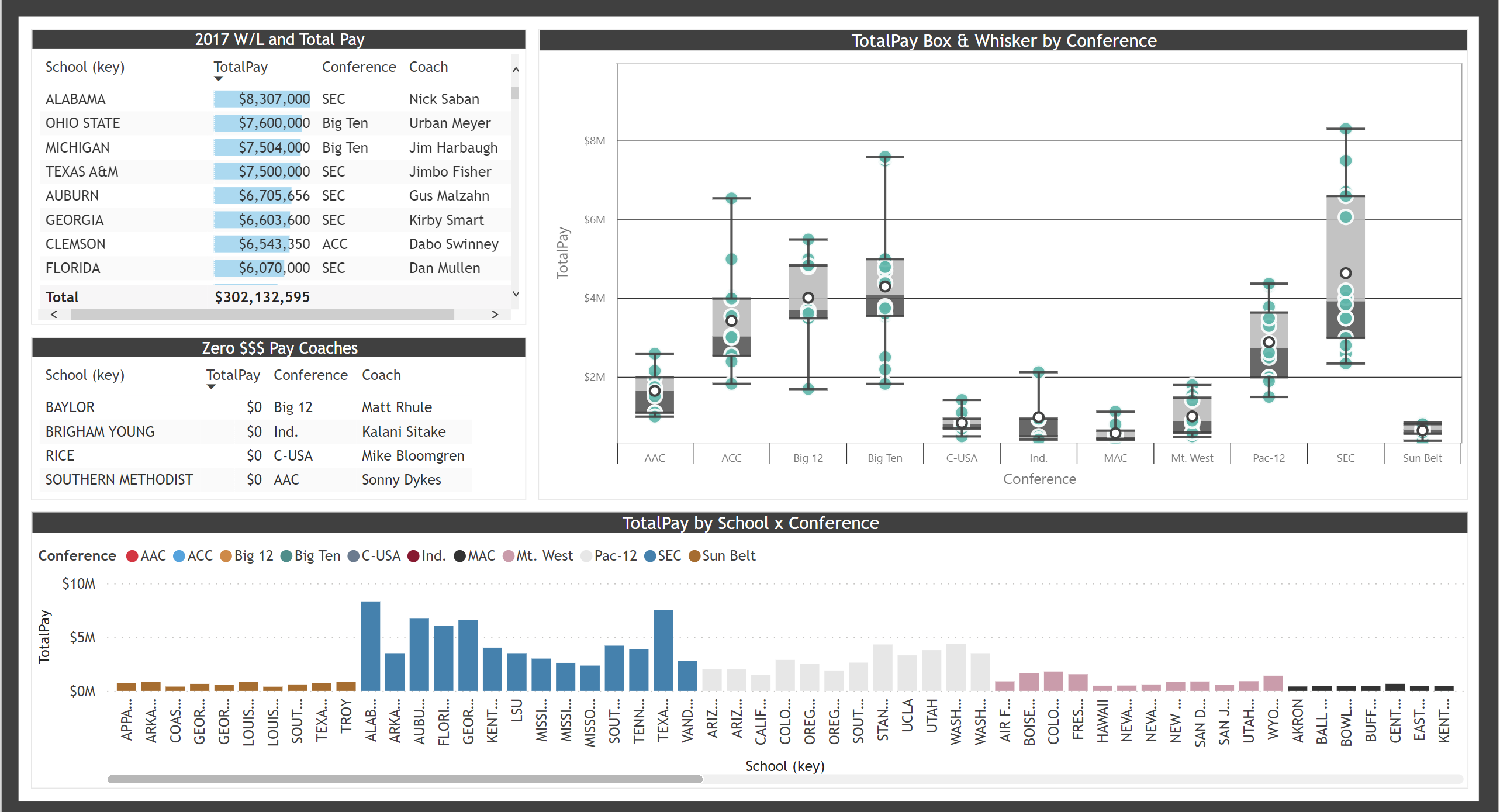
**Objective 3** – **E**xplore

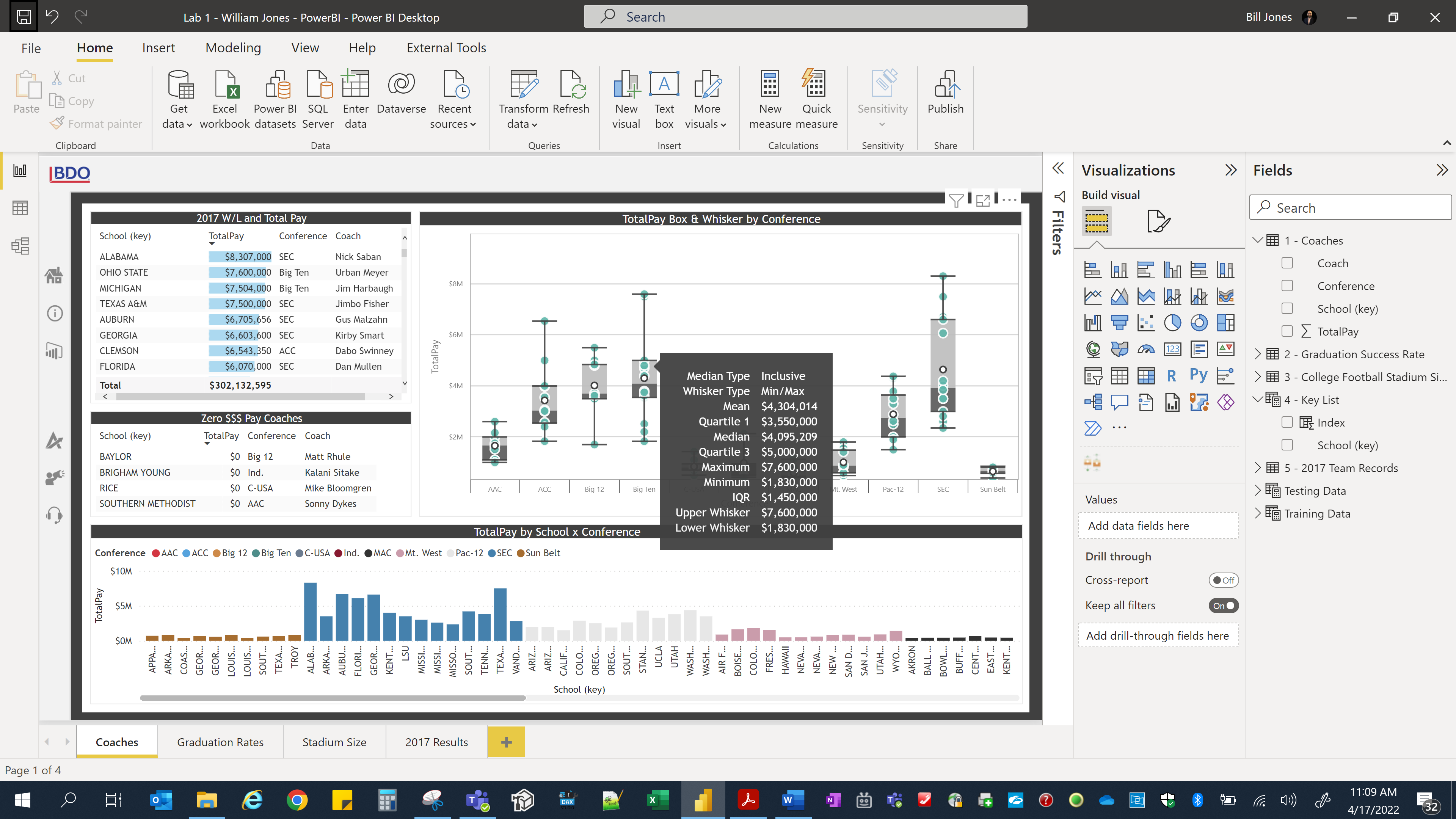
*Explore data using essential qualitative analysis techniques including descriptive statistics. Conduct an initial data analysis – exploratory data analysis – develop appropriate visualizations*

As there are 4 datasets in this analysis, I created 4 pages on a dashboard to investigate the data.

Page 1 – Coaches Data

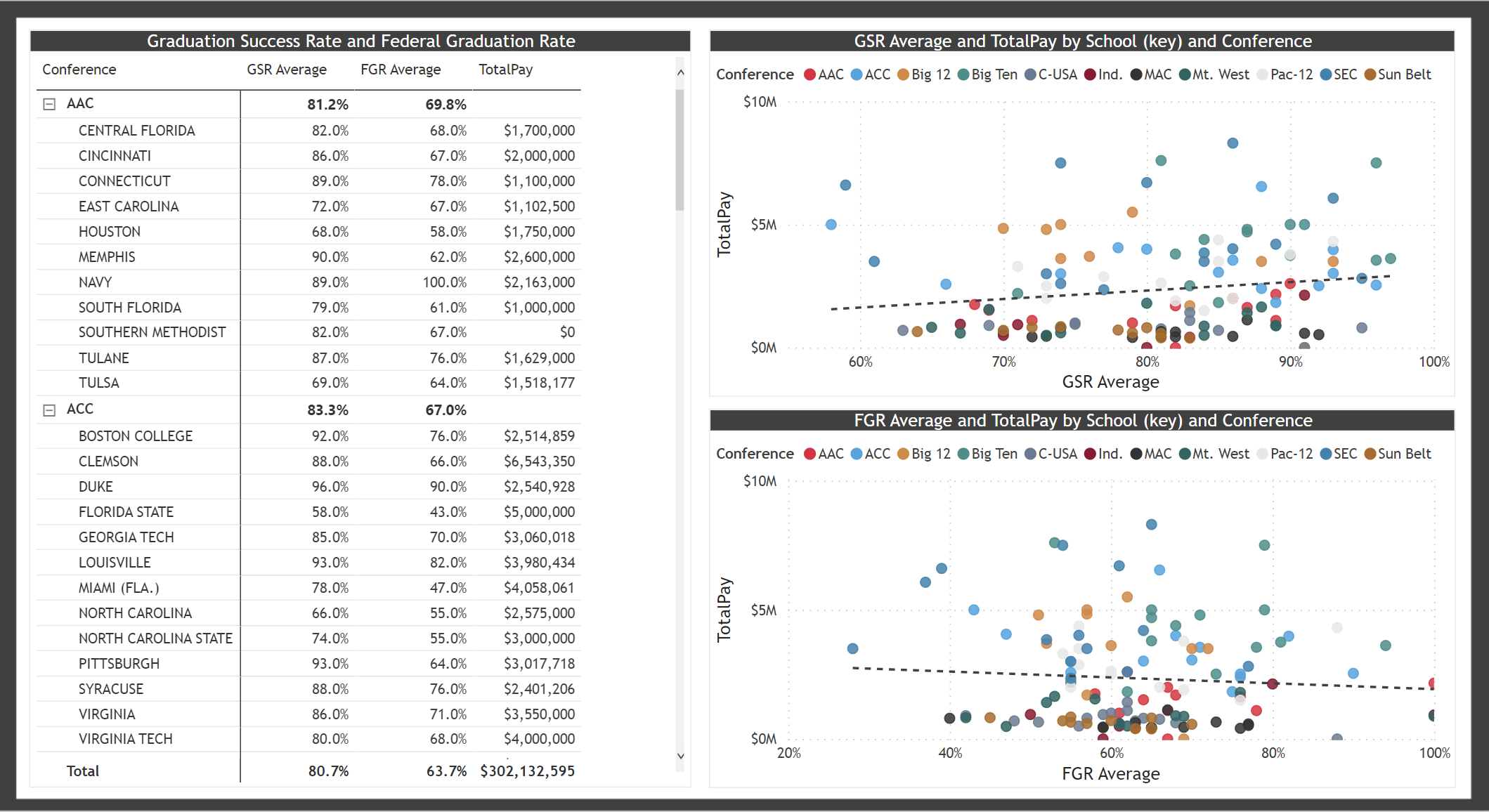
* The 2 data tables in the right are showing the coaches TotalPay data in a tabular format. This is great to include on interactive dashboards so the user can interact and see the detail. These tables help highlight the $0 TotalPay coaches which will need to be removed from the analysis to not skew the results. These $0 TotalPay coaches have been removed from all subsequent analysis.
* The 2 visualizations are helpful to show pay distribution by conference and school. The Box & Whisker visualization is quite fantastic with the ability to hover on the elements and get detailed summary results. $0 pay coaches were removed from these visualizations as evidenced by Big12 minimum ~2m and not Baylor’s $0.





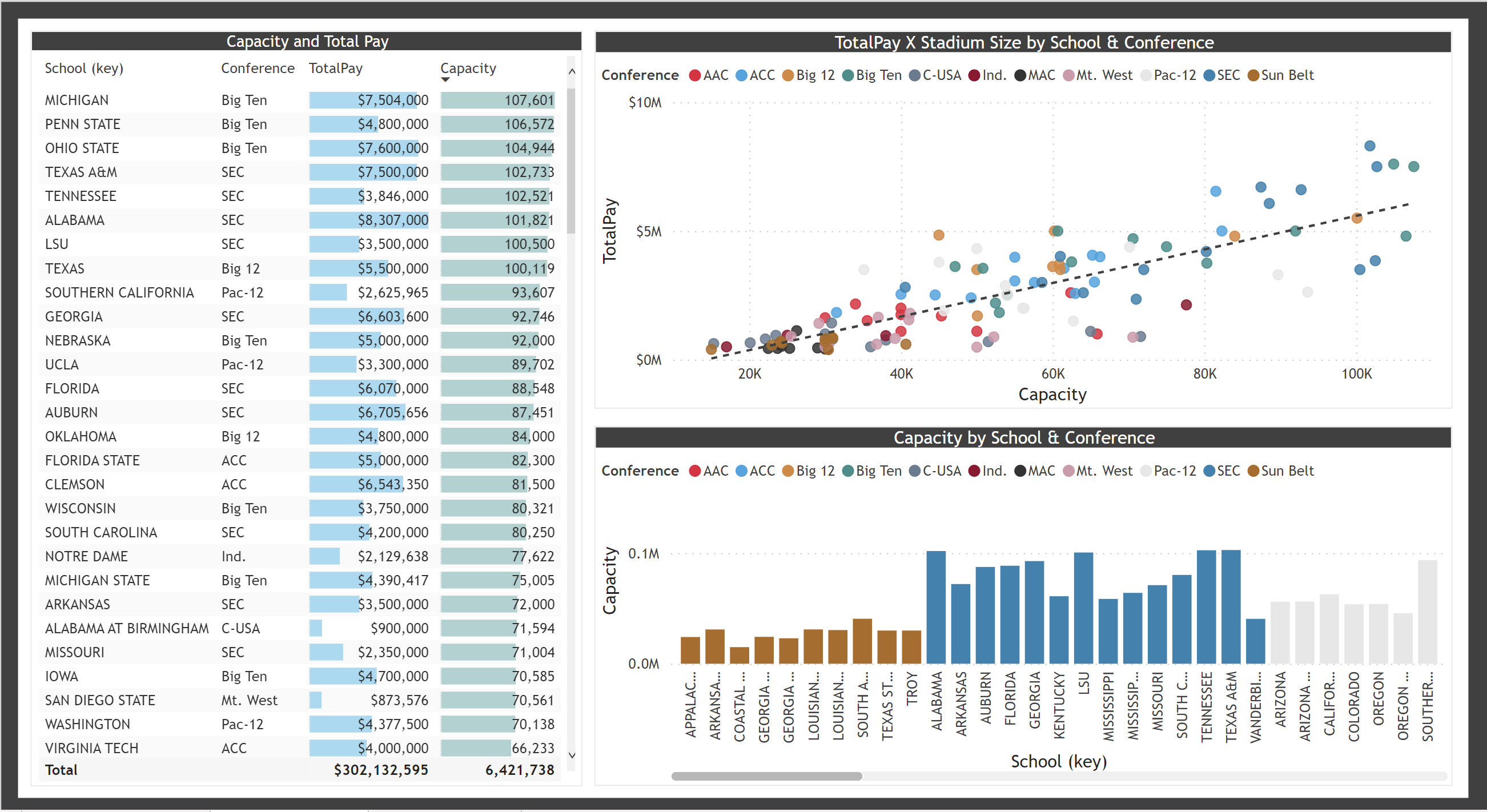
Page 2 – Graduation Rates

* Graduation rate summary table on the left showing wins and TotalPay.
* The scatterplot visualizations helps confirm the impact of Conference on pay but did not result in meaningful results for the graduation rates.



Page 3 – Stadium Size

* Capacity and TotalPay summary table on the left showing capacity and TotalPay.
* The scatterplot visualization shows TotalPay and stadium capacity have a pretty strong relationship. The bar chart visualization helps show the correlation/grouping of stadium size and conference.



Page 4 – 2017 Results

* The 2017 Results table on the left shows 2017 results detail and TotalPay by school. A column in the data included a “AP High” which was converted to a “Ranked vs Not Ranked” data field.
* The scatterplot for TotalPay vs Strength of Schedule showed a pretty strong relationship. When including the category of Ranked the clustering become clear. This relationship exits with TotalPay vs Wins and Ranked category but not as strong.
* To review the relationships between Wins, Strength of Schedule, Ranked and TotalPay, a scatterplot was created. It showed Wins and Strength of Schedule with color being Ranked and size showing TotalPay. With the clear clustering shown here, the visualization shows how these 3 variables have significant impact on TotalPay.



**Objective 4 & 5** – **M**odel & I**N**terpret

*Model relationships between data using the appropriate analytical methodologies matched to the information and the needs of clients and users. Fit a regression model with the salary as the response and the relevant predictors (i.e., you will need more than one predictor). Interpret the data, model, analysis, and findings. Communicate the results in a meaningful way.*

Using Orange Data Mining, I created a workflow to generate and test a multivariate linear regression. The linear regression was creating using the features of Capacity, Ranked, Strength of Schedule, Wins, and Conference. The results of the linear regression were tested and scored through 10-fold cross validation and 1/3 random sampling with R2 of 0.760 and 0.733, respectively. An R-Squared value of ~0.75 would indicate that ~75% of the variance of the dependent variable being studied is explained by the variance of the independent variables.

Diagram

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated Graphical user interface, text, application

Description automatically generated

The model resulted in the following Coefficients for Prediciton. These Coefficients plus a few noted assumptions produced TotalSalary predictions as noted below.

TotalPay =

Intercept ($730,070) +

Conference (1 \* $543,827) +

Table

Description automatically generatedCapacity (49,057 \* $25.7173) +

Ranked (1 \* 78,187.8) +

Strength of Schedule (6.03 \* $63,319.5) +

Wins (8 \* $74,195.2)

= $3,589,076.57

TotalPay (Big East) =

Intercept ($730,070) +

Conference (1 \* -$441,371) +

Capacity (49,057 \* $25.7173) +

Ranked (1 \* 78,187.8) +

Strength of Schedule (0.04 \* $63,319.5) +

Wins (8 \* $74,195.2)

= $2,224,594.77

TotalPay (Big Ten) =

Intercept ($730,070) +

Conference (1 \* $890,215) +

Capacity (49,057 \* $25.7173) +

Ranked (1 \* 78,187.8) +

Strength of Schedule (7.89 \* $63,319.5) +

Wins (8 \* $74,195.2)

= $4,053,238.84