2019-0703 IST 707 Data Analytics

Homework Assignment 1 (week 1)

Ryan Timbrook
NetID: RTIMBROO
Course: IST 718 Big Data Analytics
Term: Summer, 2019

Homework Assignment 1 (week 1)

Table of Contents

1 Introduction

1.1 Objective

- Combine datasets to produce meaningful analysis. Specifically, we will provide a decision maker with more than just data - we provide insights, understanding, and wisdom.

1. Obtain data and understand data structures and data elements.
2. Scrub data using scripting methods
3. Explore data using essential qualitative analysis techniques, including descriptive statistics.
4. Model relationships between data using the appropriate analytical methodologies matched to the information and the needs of clients and users.
5. Interpret the data, model, analysis, and findings, and communicate the results in a meaningful way.

1.2 Instructions

- The research question is, **how can we recommend the best salary (total compensation, minus bonus) for our next head football coach?**
- Start with the data Coaches.
- Review the data-- clean as appropriate
- Consider the base worksheet and the additional data.
    - Stadium size
    - Graduation rate:
        - Available from: NCAA Graduation Rates
        - Use the 2006 or latest cohort and include both GSR and FGR
    - Annual donations to program (if available)
- Develop an additional vector for each school using last year's record.
- Build a data frame for your analysis.
- Conduct an initial data analysis.
- Fit a regression model with the salary as the response and the relevant predictors (more than one is needed...)

1.3 Answer the following questions:

- What is the recommended salary for the Syracuse football coach?
  - Recommended Salary:
    - Current Syracuse Coach, TotalPay: $2,401,206
    - ACC:
      - **Recommend Pay: $2,727,901**
  - What would his salary be if we were still in the Big East?
    - **Predicted Pay: $2,451,775**
  - What if we went to the Big Ten?
    - **Predicted Pay: $2,520,168**
      - Currently Syracuse is in the Atlantic Coast Conference (ACC)
      - The Big East is no longer a conference and not part of this dataset.
        - A subset dataset could be considered based on the teams who belonged to that conference when Syracuse did.
- What schools did we drop from our data, and why?
  - Dropped Schools:
    - No schools were dropped from the coach's dataset.
- What effect does graduation rate have on the projected salary?
  - GSR on Projected Salary:
    - In the ACC Conference:
      - Model 2: $208
      - Model 3: $8,419
    - In the Big Ten Conference:
      - Model 2: $4,196
      - Model 3: $-1,514
    - In the Big East Conference:
      - Model 2: $2,051
      - Model 3: $6,871
    - Overall of the coach's dataset:
      - Model 2: $6,398
      - Model 3: $8,956
- How good is our model?
  - Overall Model 4 performed the best on each of the test scenarios. For all records:
    - M4: Proportion of Test Set Variance Accounted for: 0.485
    - M4: Most significant attribute: 'Score' with value: $67,279.0

    For the ACC Conference:
    - M4: Proportion of Test Set Variance Accounted for: 0.858
    - M4: Most significant attribute: 'WLRatio' with value: $214,497.0

    For the Big Ten Conference:
    - M4: Proportion of Test Set Variance Accounted for: 0.052
    - M4: Most significant attribute: 'WLRatio' with value: $73,876.0

    For the Big East Conference:
    - M4: Proportion of Test Set Variance Accounted for: 0.985
    - M4: Most significant attribute: 'Score' with value: $214,301.0

*Inconsistencies in the model accuracy over the difference subdivisions is mostlikely due to unaccounted for outliers.

- What is the single biggest impact on salary size?
  - Biggest Impact on Salary:
    - In the ACC Conference grouping, **'WLRatio' with a net increase of $214,497.**

Feature Sets:

- Model 2: BonusPaid + StadSize + GSR + SeatRank + GSRank + W + L + WLRatio + OffenceScore + DefenseScore + Score + PointsPerGame
- Model 3: WLRatio + StadSize + SeatRank + GSR + GSRank + Score + PointsPerGame
- Model 4: Score + WLRatio + StadSize

Best feature Set Performance: Model 4

- Score + WLRatio + StadSize
  * Best Feature:
    *Score

2 Analysis and Models

2.1 About the Data

Coaches Dataset:

Coaches Dataframe Shape: (118, 23)
Coaches Dataframe, Number of records: 118
Coaches Dataframe Size: 2714

NCAA Graduation Rate Datasets:

- Available from: NCAA Research Data

NCAA GSR Dataframe Shape: (5403, 12)
NCAA GSR Dataframe, Number of records: 5403
NCAA GSR Dataframe Size: 64836

NCAA Stadium Capacity Dataset:

- Available from GithHub: gboeing/data-visualization

NCAA Stadium Capacity Dataframe Shape: (253, 13)
NCAA Stadium Capacity, Number of records: 253
NCAA Stadium Capacity: 3289

*These three datasets were linked by School/Team name, and conference

2.1.1 *Data Exploration & Cleaning*

The coach's dataset needed heavy cleaning and transformation.
No records were removed from the dataset during the cleaning processes.

The first step was identifying where the missing values where.

- Total Count of NaN in Coaches Dataset: 187
- Rows that contain NaN values in Coaches Dataset: 59

The second step was understanding the attributes that had the missing values. It was found that for a good majority of those items, they were of the same record.

- Pay Type Attributes:
  - SchoolPay: [11, 15, 91, 94, 95]
  - TotalPay: [11, 15, 91, 94, 95]

- Scoring Attributes:
  - W: [14, 18, 23, 48, 49, 56, 57, 61, 82, 90, 91]
  - L: [14, 18, 23, 48, 49, 56, 57, 61, 82, 90, 91]
  - WLRatio: [14, 18, 23, 48, 49, 56, 57, 61, 82, 90, 91]
  - OffenceScore: [14, 18, 23, 48, 49, 56, 57, 61, 82, 90, 91]
  - DefenseScore: [14, 18, 23, 48, 49, 56, 57, 61, 82, 90, 91]
  - Score: [14, 18, 23, 48, 49, 56, 57, 61, 82, 90, 91]
  - PointsPerGame: [14, 18, 23, 48, 49, 56, 57, 61, 82, 90, 91]

For the scoring type attributes, missing values were replaced by their conference's median value for that attribute.

For the Bonus & BonusPaid attributes, those that were missing were given a value of 0. The PayPlusBonus2016 values were replaced with the TotalPay value of that record.

Investigation of the three primary Conferences:
ACC, Big Ten and Big East, showed that these conferences do not have a normally distributed salary range. This is show below in the TotalPay figures.

Note: Big East is no longer a football conference. Teams from when Syracuse was a member of that conference have been aggregated together to create a mock conference scenario.**

2.1.2 *Data Transformations*

New data frames were created with linking attributes for data lookup, and the creation of a mock Big East data set formed in-order to answer the salary question about if Syracuse was back in the Big East. Since the Big East Football conference doesn't exist any longer, a mocked dataset was created based on those teams who belonged to it during the 2012 session with Syracuse.

2.1.3 *Exploratory Data Visualizations*

Figure: ACC School Pay distribution

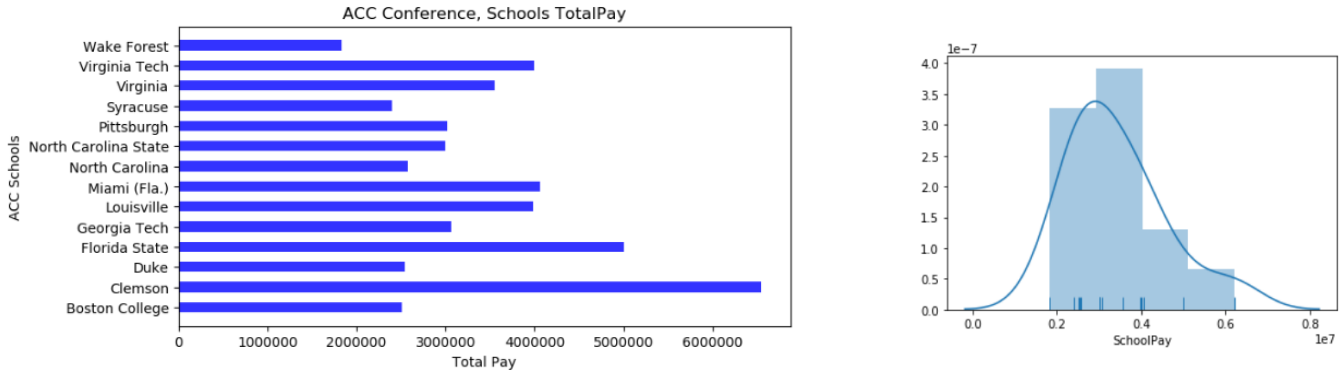
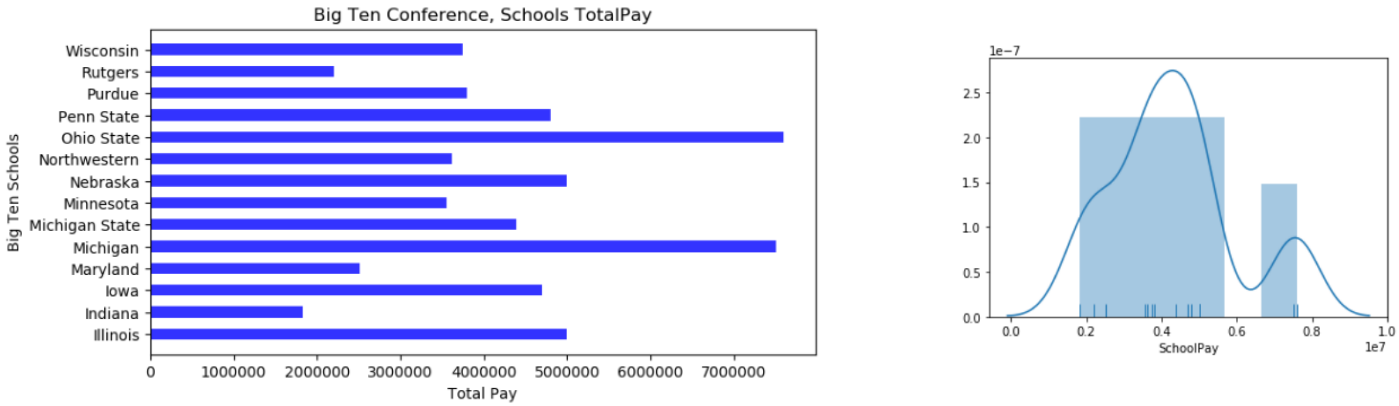


Figure: Big Ten School Pay distribution



Homework Assignment 1 (week 1)

Figure: Big East School Pay distribution

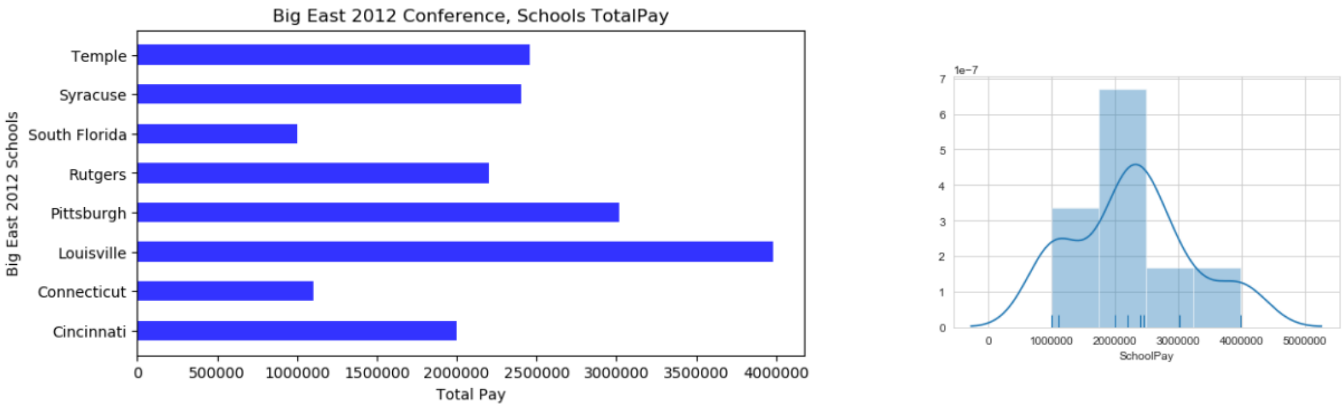


Figure: TotalPay distribution over all Conferences

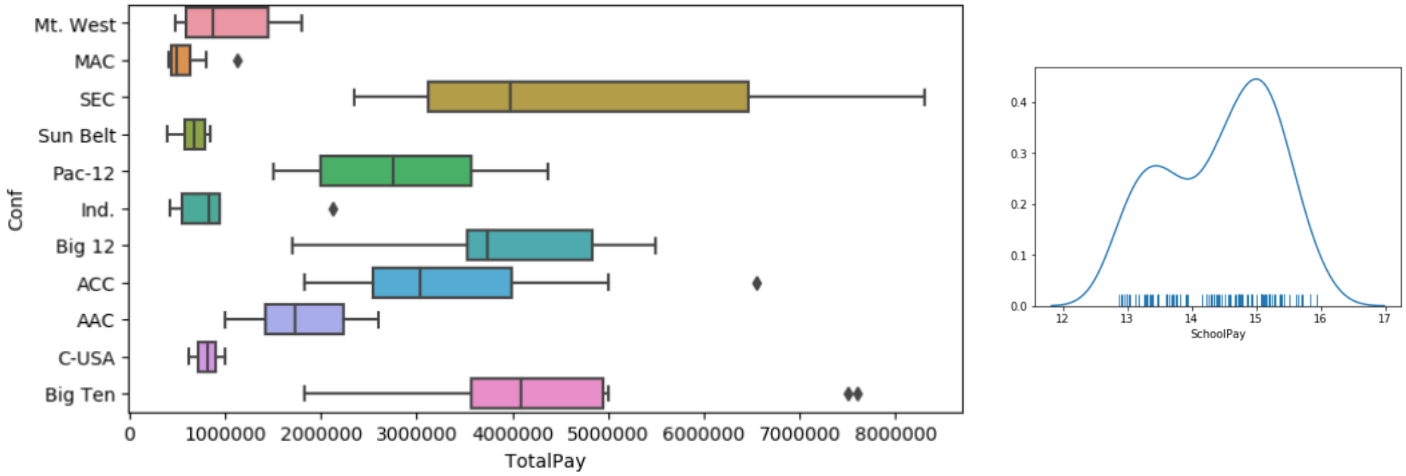
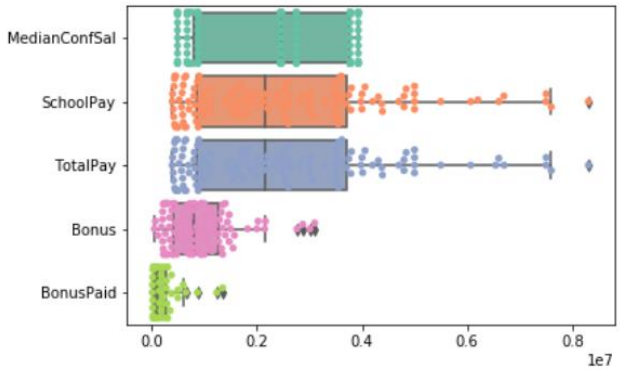


Figure: Distribution of Pay Types



Homework Assignment 1 (week 1)

Figure: Figure: Z-Scaled pair plot of ['WLRatio','GSRank','StadSize'] on TotalPay

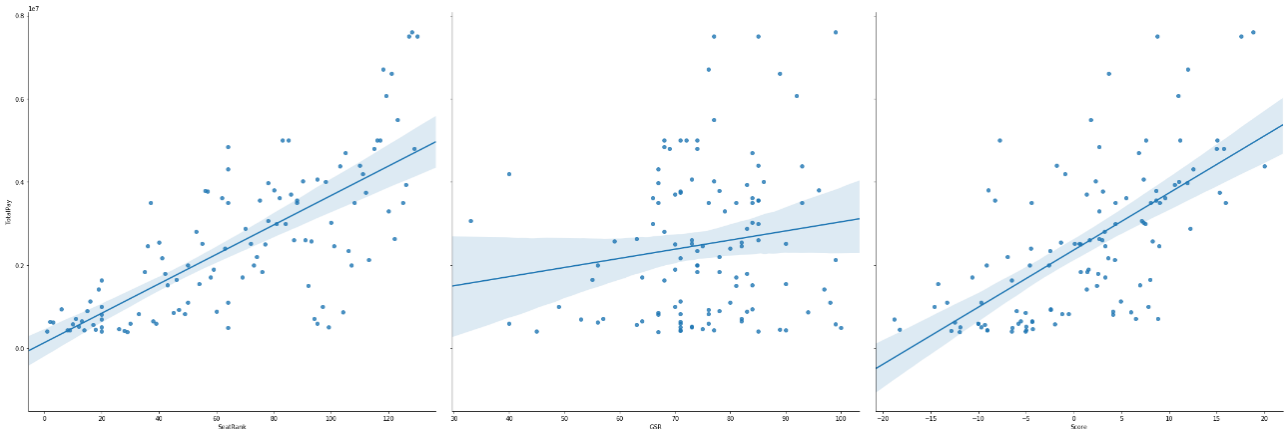


Figure: Z-Scaled pair plot of ['SeatRank','GSR','Score'] on TotalPay

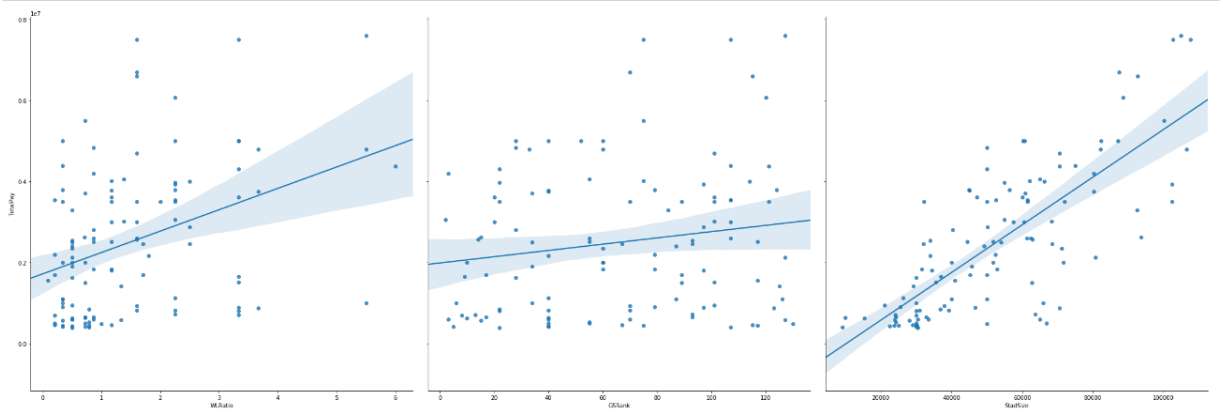


Figure: Correlation Heatmap, Z-score scaled attributes

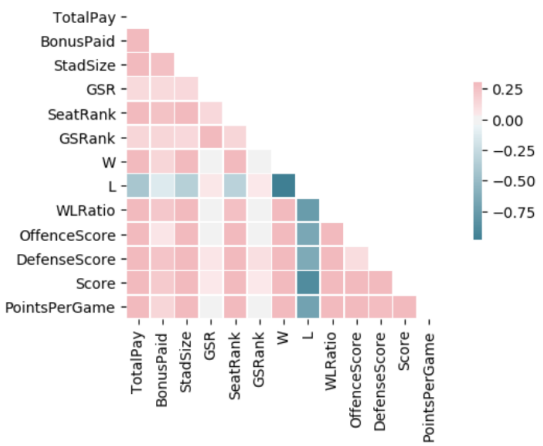
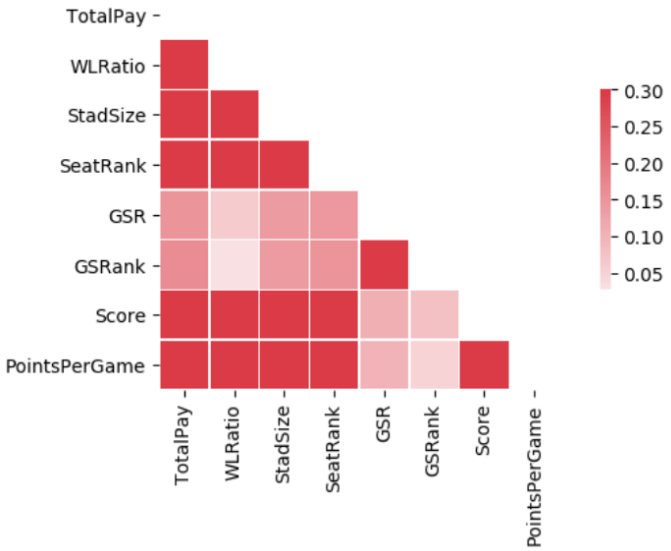


Figure: Correlation Heatmap, Non-Z-score



2.2 Models

2.2.1 *Model 1 Details*

Single Regression - First Run
Ordinary Least Squares Regression Feature:

- Dependent Variable (Response Variable: TotalPay
- Independent Variable (Explanatory Variable): WLRatio

M1: Using z-score normalization had a positive effect on model performance:

- Test Set Variance Accounted for raised from 0.06 to: 0.278
- R-squared reduced from 0.242 to: 0.093
- F-statistic changed from 26.17 to: 8.170
- WLRatio coef, P-Value changed from 0.000 to: 0.005

Parameters: 'TotalPay ~ WLRatio'

2.2.2 *Model 2, 3, 4 Details*

Multiple Regression
Feature Set:

- M2: BonusPaid + StadSize + GSR + SeatRank + GSRank + W + L + WLRatio + OffenceScore + DefenseScore + Score + PointsPerGame
- M3: WLRatio + StadSize + SeatRank + GSR + GSRank + Score + PointsPerGame
- M4: WLRatio + StadSize + GSRank

Best feature Set:

- Score + WLRatio + StadSize
    - Best Feature:
        - Score

Training-Test sampling taken at 1000x with replacement each to ensure there was ample data for the training/test splits.

2.2.3 *Performance*

2.2.4 *z-score normalization effects on M2:*

- Test Set Variance Accounted from: 0.445 to: 0.748
- R-squared changed from: 0.836 to: 0.713
- Adj. R-squared changed from: 0.809 to: 0.664
- F-statistic changed from: 30.23 to: 14.32
    - Coef P-values changed from:
        - WLRatio from: 0.514 to: 0.123
        - BonusPaid from: 0.010 to: 0.007
        - StadSize from: 0.000 to: 0.040
        - GSR from: 0.490 to: 0.756
        - SeatRank from: 0.078 to: 0.470

- GSRank from: 0.320 to: 0.979
- W from: 0.615 to: 0.572
- L from: 0.394 to: 0.456
- OffenceScore: 0.378 to: 0.110
- DefenseScore: 0.675 to: 0.349
- Score: 0.354 to: 0.117
- PointsPerGame: 0.015 to: 0.022

2.2.5 *z-score normalization effects on M3:*

- Test Set Variance Accounted from: 0.395 to: 0.815
- R-squared changed from: 0.764 to: 0.644
- Adj. R-squared changed from: 0.743 to: 0.611
- F-statistic changed from: 35.23 to: 19.15
  - Coef P-values changed from:
    - WLRatio from: 0.753 to: 0.028
    - StadSize from: 0.000 to: 0.021
    - GSR from: 0.510 to: 0.589
    - SeatRank from: 0.186 to: 0.460
    - GSRank from: 0.138 to: 0.488
    - Score: 0.020 to: 0.001
    - PointsPerGame: 0.755 to: 0.726

  - Overall Model 4 performed the best on each of the test scenarios. For all records:
    - M4: Proportion of Test Set Variance Accounted for: 0.485
    - M4: Most significant attribute: 'Score' with value: $67,279.0

    For the ACC Conference:
    - M4: Proportion of Test Set Variance Accounted for: 0.858
    - M4: Most significant attribute: 'WLRatio' with value: $214,497.0

    For the Big Ten Conference:
    - M4: Proportion of Test Set Variance Accounted for: 0.052
    - M4: Most significant attribute: 'WLRatio' with value: $73,876.0

    For the Big East Conference:
    - M4: Proportion of Test Set Variance Accounted for: 0.985
    - M4: Most significant attribute: 'Score' with value: $214,301.0

TotalPay Predictions:

Syracuse Coach, TotalPay: $2,401,206
All:
Predicted Pay: $2,090,379
ACC:
Predicted Pay: $2,727,901
Big Ten:
Predicted Pay: $2,520,168
Big East:
Predicted Pay: $2,451,775

Homework Assignment 1 (week 1)