Shan Ali, Callie Gilmore, Jocelyne Walker

College Football Game Attendance

Description of Project Goals

College football games have been a tradition for students, alumni, and fans for decades. However, in recent years, attendance has decreased at NCAA games. According to a Fall 2019 NPR interview, from 2004 to 2018, average stadium capacity has fallen 2400 seats, and average attendance has fallen nearly 3500 seats.[1]

Rising ticket prices, crowded streets, and the exposure to the hot sun or cold winters deter more fans from cheering their teams on in person. Instead, the at-home experience seems much more pleasant: turning on the TV, sitting on your couch, and drinking beer from your own fridge. The University of Texas at Austin's Athletic Director Chris del Conte has taken measures to turn game day into an *experience* where fans want to attend in person.[2] Live music, tailgating, and reduced concession stand markups all contributed to increasing attendance throughout the 2019 season. So how can the NCAA and universities ensure that fan attendance doesn't continue to fall?

We'll use college football attendance data to answer two main questions:

- Is the decreasing attendance trend described in the NPR article proved in our data after adjusting for other predictors?
- What factors are most significant to determine game attendance?

Dataset Description

We used a Kaggle college football game dataset scraped from Wikipedia. It contains data from 18 different football seasons, from 2000 to 2018, and 63 different teams across each of the 14 Division 1 conferences. It is important to note that the data does not come from every Division 1 team but a sample of 63 teams where each conference is represented at least once. Each of the 6,672 rows represents a unique game played. For each game, the dataset provides 25 insights including attendance, weather, date, rankings, score, site, and stadium capacity (Table 1). For our goals, we decided to focus on exploring and predicting gameday attendance.

Cleaning the Data

In order to analyze the data, we needed to clean the raw data. The raw dataset began with 25 predictors and, after cleaning, had 32 predictors. We began by dropping all NaN entries.

Predictors *opponent*, *result*, and *site* were split into multiple new categories to expand the embedded information to use as new predictors. For example, the *site* predictor was expanded to add a *city*, *state*, *region*, and *special*

[1] Attendance Drops For College Football

[2] Game day reimagined: Texas style

game predictor as well as remove the redundant rank and shorten the site to only the location string. Predictors *tailgating*, *Win/Loss*, *New Coach* were converted into dummies to facilitate analysis and reduce redundant information. For example, the *tailgating* predictor was converted into 0, 1 dummies (True = 1) to better encode the data for analysis. The date, time, and year predictors were consolidated into a single *DateTime* object to facilitate analysis. Finally, we dropped all redundant predictors to clean the dataset and reduce overfitting.

Exploratory Analysis
To understand the data better, we ran several visualizations. One of interest was a pie chart displaying the ratio of teams from each conference displayed in the dataset. From Figure 2, half the dataset is composed of teams from 4 conferences: WAC, Mid-American, ACC and Big 12. The other half of the data is made up of the remaining 10 conferences.

We displayed different visual analyses related to attendance to see any insights or trends. Conferences with the highest average attendance are the SEC, Big-10, and Independent Teams (Figure 3). Teams with the highest average attendance are Penn State (Big-10), Alabama (SEC), and Texas A&M (now SEC) (Figure 4). Conferences, teams, and regions could potentially have a high impact as some teams and conferences are more popular or larger than others and will therefore have higher attendance. These results also make sense when comparing *stadium capacity*. Teams with larger stadium capacities will also have higher attendance due to availability of seats. Similarly, the conferences with the highest stadium capacities are the SEC, Big-10 and Independent Teams (Figure 5), and the teams with largest stadium capacities are Penn State, Texas A&M and Alabama (Figure 6).

We extracted the state value from the site column and created a heat map of mean attendance by state (Figure 7). The states with the highest average attendance are Pennsylvania, Nebraska, and Oklahoma, consistent with the top 10 teams shown in Figure 4.

Solution
In order to predict college football game attendance, we thought about all the potential predictor variables in our dataset. Some X variables included in the dataset that had to be removed in order to build an attendance model:

- *Team, Opponent, Site,* and *DateTime* were removed because these variables have too many distinct values. There are 63 distinct teams, 343 distinct opposing teams, 252 distinct sites, and 4000 dates.
- *Win, Team Score, Opponent Score,* and *OT* were removed because these variables occur *after* the game and thus do not contribute to game attendance. They could be *response* variables in other models.

From our exploratory analysis, we had an intuition of which predictors might be most important to gameday attendance.

First, we ran an Ordinary Least Squares model using Python (Figure 8). We split the data into training and testing sets, with the first 80% of the data (2000-2015) used to train and the last 20% (2016-2018) used to cross validate. The metric used to choose our model was the out-of-sample Root Mean Squared Error (RMSE). Predictors included in the model were *Conference, Current Losses, isRanked, OppisRanked, and Tailgating.* The most statistically significant predictors were **Conference** and **Tailgating**, though all of the predictors were important. This makes sense as the conference speaks to the popularity of teams in that group. From the coefficients, we see that, holding other predictors constant, the SEC conference has an estimated lift of 49,610 fans and Big-12 has an estimated lift of 45,240 fans. Smaller conferences like Mid-American have estimated lifts of only 18,370 fans. Tailgating at the game has an estimated coefficient of 27,600 fans. This makes sense as fans are more likely to attend games that are full-day events, with parties before the game.

The mean of attendance is 45,303 fans with a standard deviation of 25,184 fans. The in-sample RMSE of this model was 12,628 fans, with an out-of-sample RMSE of **13,692** fans, an improvement over the standard deviation. This model was largely based on intuition from exploratory analysis. Next, we statistically calculated which predictors are most important and checked if there was anything surprising.

We pulled the cleaned College Football DataFrame out of Python and into R to use additional predictive models. As a baseline, we ran a multiple linear regression predicting attendance using **all** predictors, a total of **73** once all factors were created for each conference and state (Figure 9). This model showed significant improvement over the initial model. However, there was evidence of overfitting as the in-sample RMSE was close to 0, a large difference compared to the out-of-sample RMSE of **7500** fans.

To reduce overfitting and improve the model, we also ran a Lasso regression to obtain a subset of the most important predictors (Figure 10). Sixteen predictors were removed from the original model, resulting in **57** factors. This model predicted the in-sample and out-of-sample attendance much more closely, with an in-sample RMSE of 7260 and an out-of-sample RMSE of **7490** fans.

From our Lasso model, the most important predictors for Attendance are:

- In State, **California**, **Tennessee**, and **Georgia** all *lower* game attendance. As state was not included in the factor of our original model, this shows that certain states do tend to have lower in-person gameday attendance.

- **Tailgating**, **Big-10** conference, **SEC** conference, and **Time** of day all contribute to higher game attendance. The presence of tailgating and Big-10 and SEC conferences having highest attendance is consistent with previous models and our exploratory analysis. It also makes sense that Time of day contributes to increased attendance as (1) more fans are likely to attend games later in the day, and (2) the most popular games usually occur during the prime-time evening time slots.

Insights

- **Small downward attendance trend is present in the regression models.** The coefficient for Year is -68 in the multiple regression and -38 in the Lasso model. While this coefficient is small, it is statistically significant and verifies the slight downward trend in attendance, even after adjusting for other predictor variables.
- **The most important factors for in-person attendance are the presence of Tailgating and the Conference teams play in**. Thus, increasing the visibility of less prestigious conferences by playing ranked teams and in-person events draw fans to the game (Figure 10).
- **Attendance ranges differ significantly in large ranked schools versus smaller schools.** UTEP, a small CUSA conference school, has an attendance range with a min of 9690 fans and a max of 53415, a range of over **43,000 fans** (Figure 11). Oklahoma has a range of only **13,000 fans** with a min of 74,432 and a max of 88,308. This suggests the need to model individual conferences' and schools' attendance separately for more granular insights.
- **Texas A&M's move to the SEC was smart for increasing game attendance** (Figure 12). Based on the Lasso model, SEC attendance on average is higher than Big-12 attendance, a lift of 7065 fans versus 5691 fans, other variables held constant. On average, SEC teams *do* have higher attendance than non-SEC teams.

In addition to predicting gameday attendance, schools must also **price tickets** based on attendance. Having average ticket price data along with this expected demand would allow schools to maximize revenue while ensuring their stadiums are full. This model could also be improved by considering the impact of both **stadium upgrades** and **stadium expansions** on increasing game day attendance. While stadium expansions are naturally included in the model due to the stadium capacity predictor, the impact of stadium upgrades like Texas football's $179M current stadium upgrade could be statistically modeled based on similar schools' upgrades and resulting ticket price, attendance, and revenue increases.

Appendix

Figure 1: Attendance, capacity, and fill rate are falling over the last 18 years

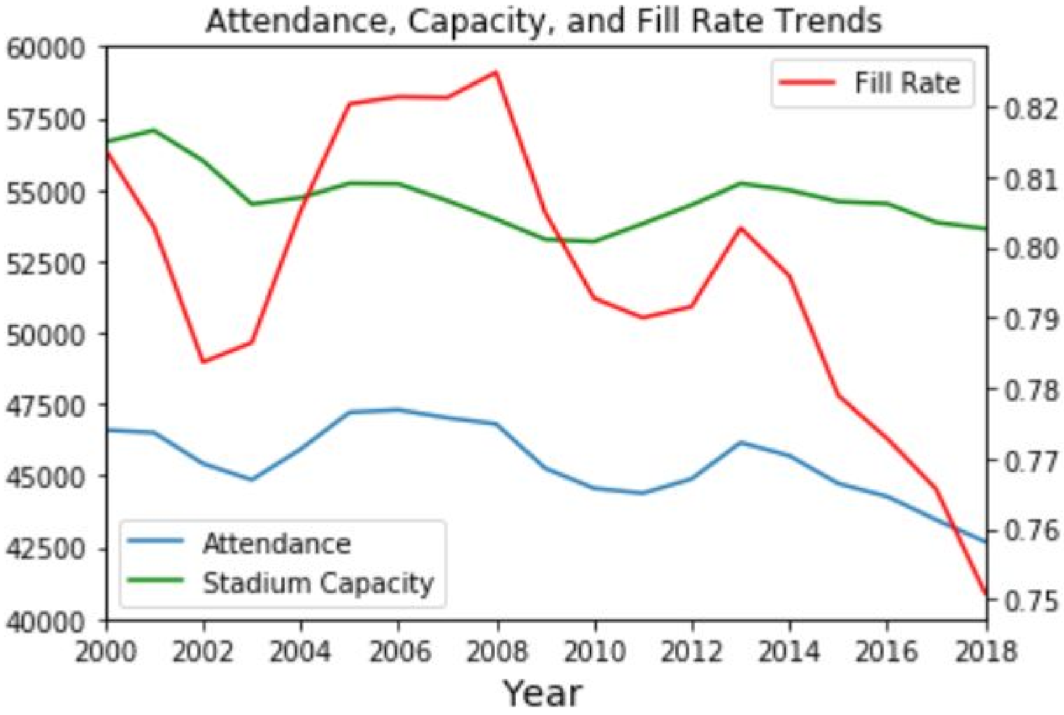


Figure 2: Four conferences (Big-12, ACC, Mid-American, and WAC) represent over 50% of the teams in the sample

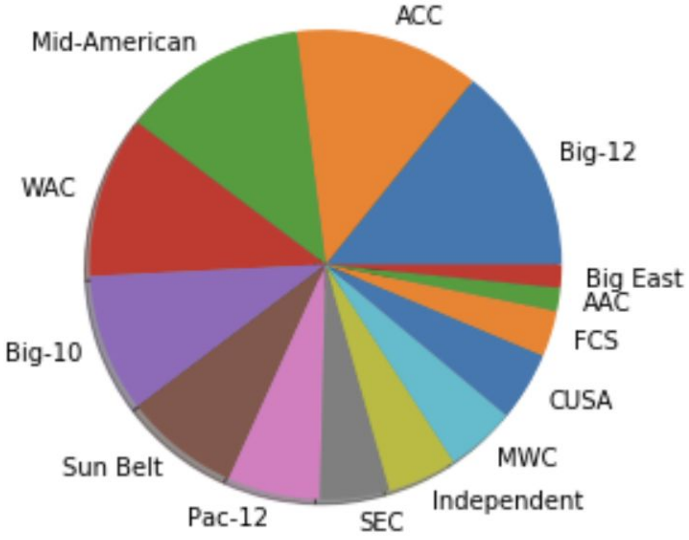


Figure 3: SEC, Big-10, and Independent teams have highest on average attendance

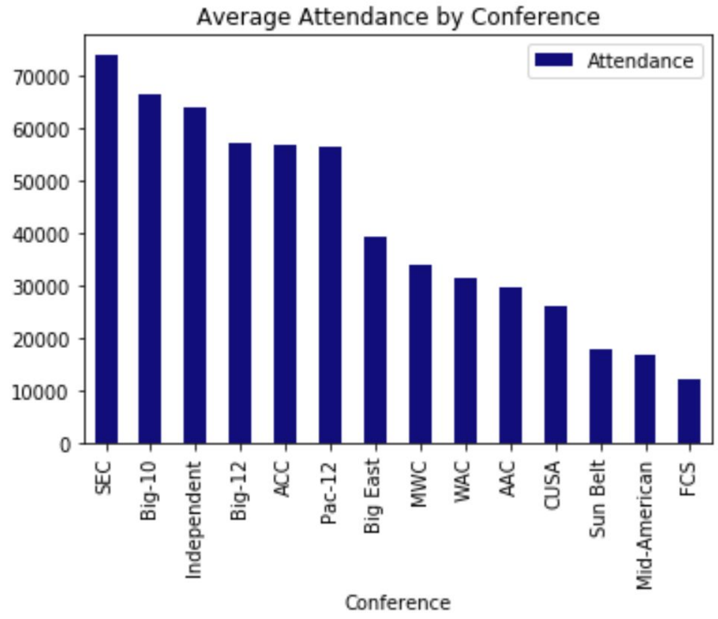


Figure 4: Penn State, Alabama, and Texas A&M have highest average game day attendance

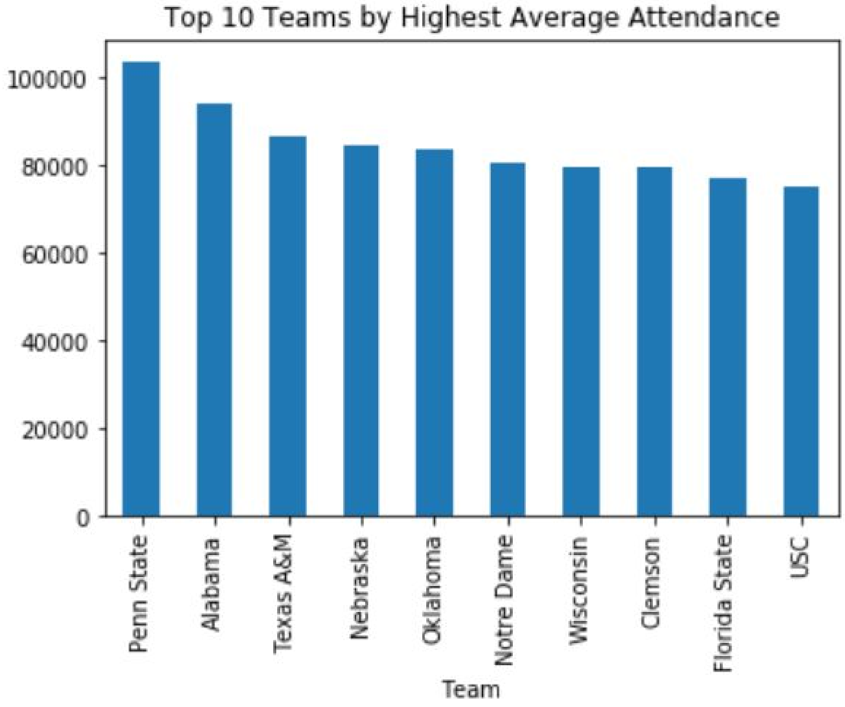


Figure 5: The Conferences with the highest stadium capacities are SEC, Big-10, and Independent

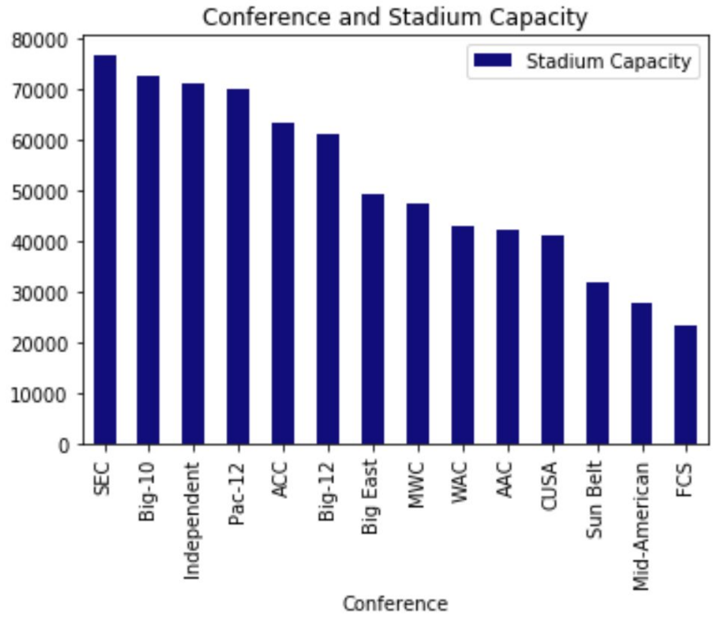


Figure 6: The Top 10 Teams by Maximum of Stadium Capacity (accounting for expansions) are Penn State, Texas A&M, and Alabama

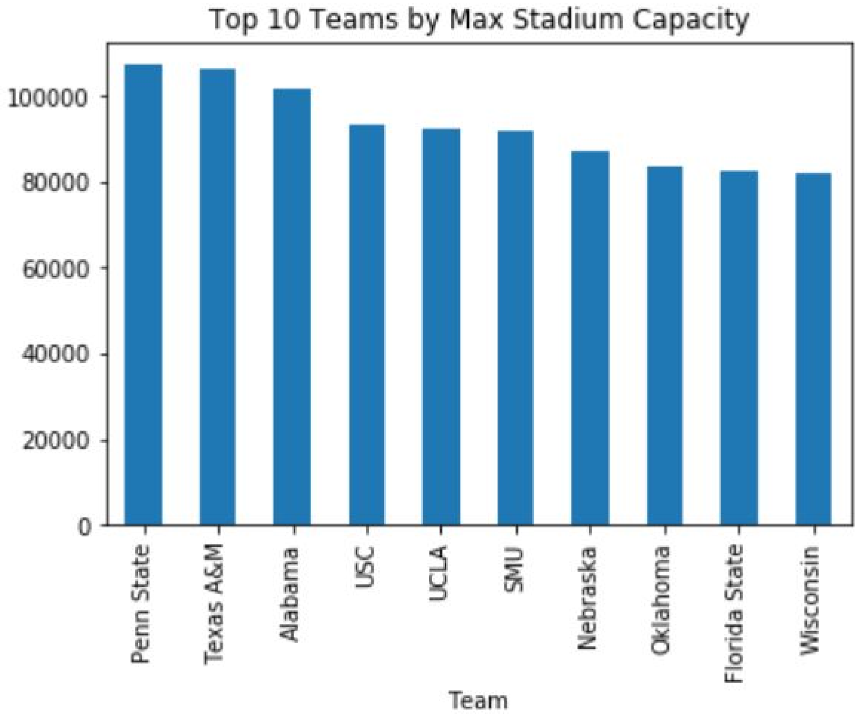


Figure 7: Heatmap of Average Attendance shows spikes in Pennsylvania, Oklahoma, and Nebraska

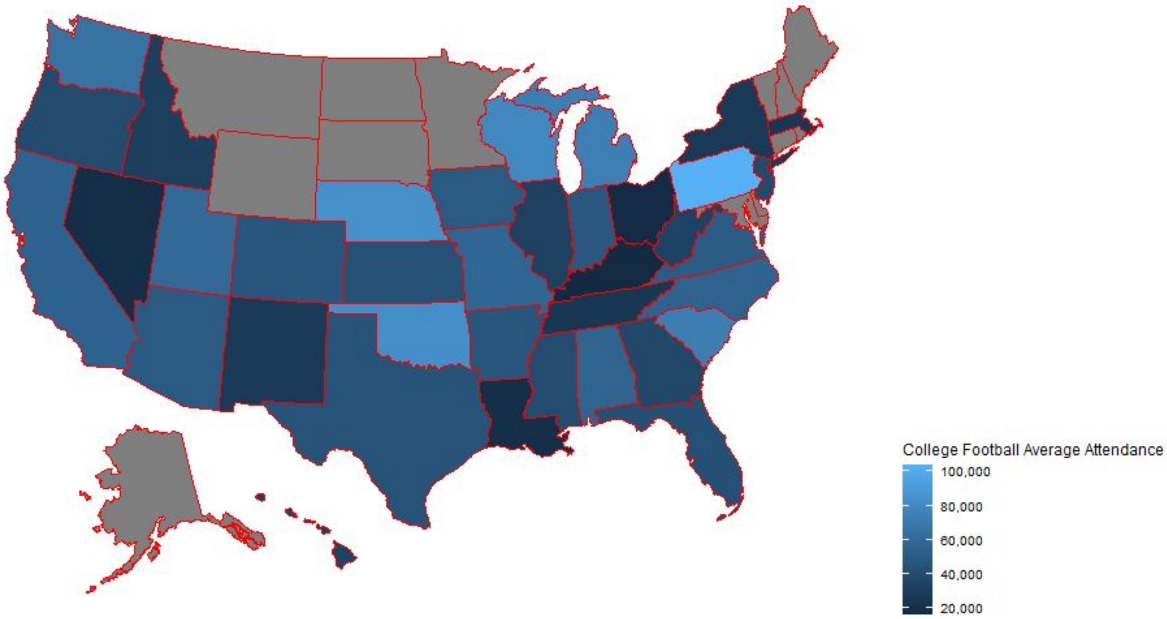
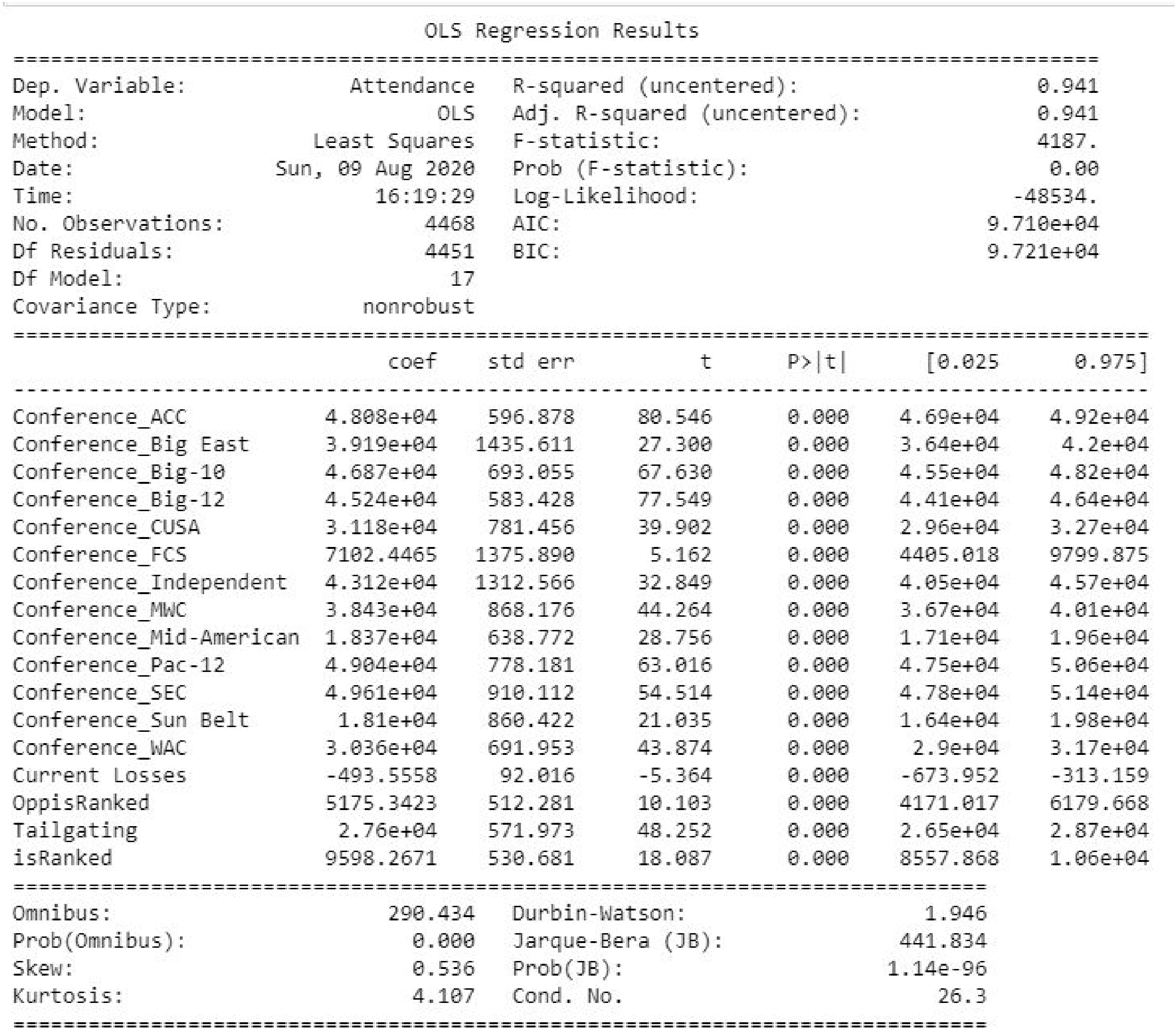
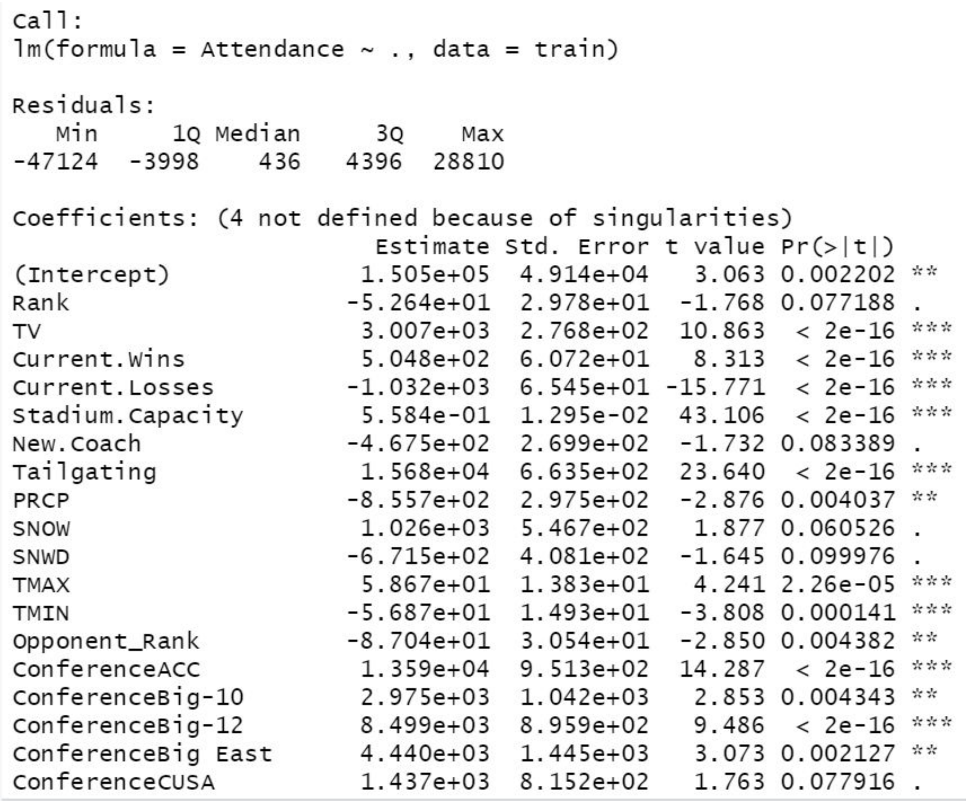


Figure 8: Python OLS Regression Results

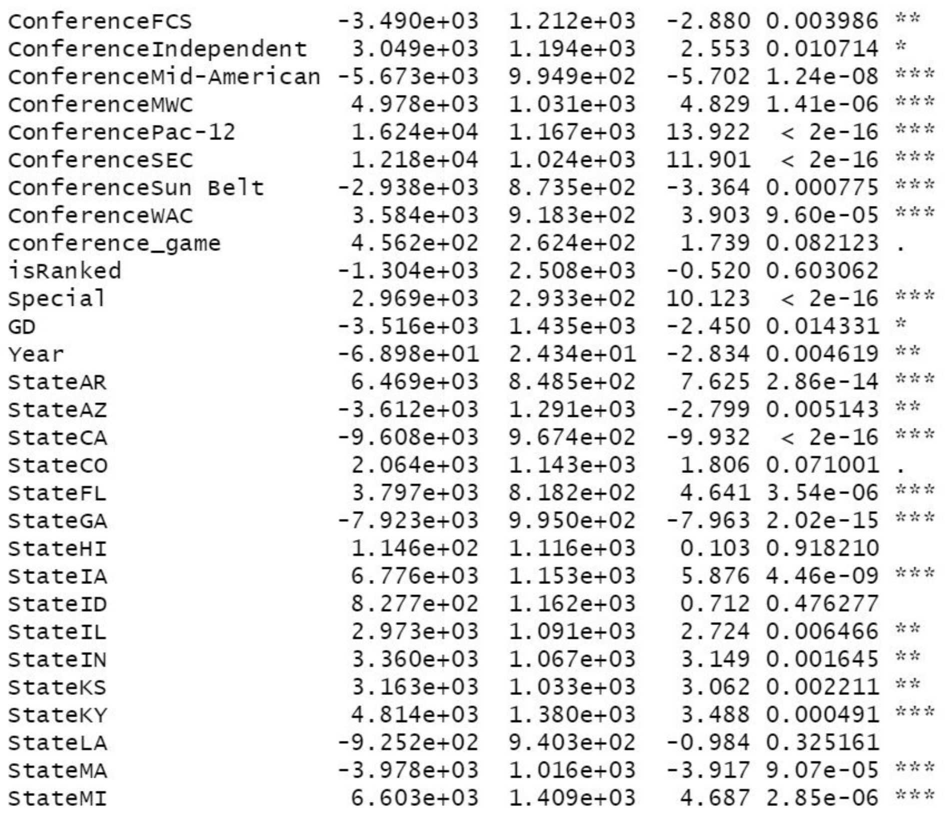


```
                                 OLS Regression Results                                
=======================================================================================
Dep. Variable:             Attendance   R-squared (uncentered):                   0.941
Model:                            OLS   Adj. R-squared (uncentered):              0.941
Method:                 Least Squares   F-statistic:                              4187.
Date:                Sun, 09 Aug 2020   Prob (F-statistic):                        0.00
Time:                        16:19:29   Log-Likelihood:                         -48534.
No. Observations:                4468   AIC:                                  9.710e+04
Df Residuals:                    4451   BIC:                                  9.721e+04
Df Model:                          17                                                  
Covariance Type:            nonrobust                                                  
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
Conference_ACC           4.808e+04    596.878     80.546      0.000    4.69e+04    4.92e+04
Conference_Big East      3.919e+04   1435.611     27.300      0.000    3.64e+04     4.2e+04
Conference_Big-10        4.687e+04    693.055     67.630      0.000    4.55e+04    4.82e+04
Conference_Big-12        4.524e+04    583.428     77.549      0.000    4.41e+04    4.64e+04
Conference_CUSA          3.118e+04    781.456     39.902      0.000    2.96e+04    3.27e+04
Conference_FCS           7102.4465   1375.890      5.162      0.000    4405.018    9799.875
Conference_Independent   4.312e+04   1312.566     32.849      0.000    4.05e+04    4.57e+04
Conference_MWC           3.843e+04    868.176     44.264      0.000    3.67e+04    4.01e+04
Conference_Mid-American  1.837e+04    638.772     28.756      0.000    1.71e+04    1.96e+04
Conference_Pac-12        4.904e+04    778.181     63.016      0.000    4.75e+04    5.06e+04
Conference_SEC           4.961e+04    910.112     54.514      0.000    4.78e+04    5.14e+04
Conference_Sun Belt       1.81e+04    860.422     21.035      0.000    1.64e+04    1.98e+04
Conference_WAC           3.036e+04    691.953     43.874      0.000     2.9e+04    3.17e+04
Current Losses           -493.5558     92.016     -5.364      0.000    -673.952    -313.159
OppisRanked              5175.3423    512.281     10.103      0.000    4171.017    6179.668
Tailgating                2.76e+04    571.973     48.252      0.000    2.65e+04    2.87e+04
isRanked                 9598.2671    530.681     18.087      0.000    8557.868    1.06e+04
==============================================================================
Omnibus:                      290.434   Durbin-Watson:                   1.946
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              441.834
Skew:                           0.536   Prob(JB):                     1.14e-96
Kurtosis:                       4.107   Cond. No.                         26.3
==============================================================================
```

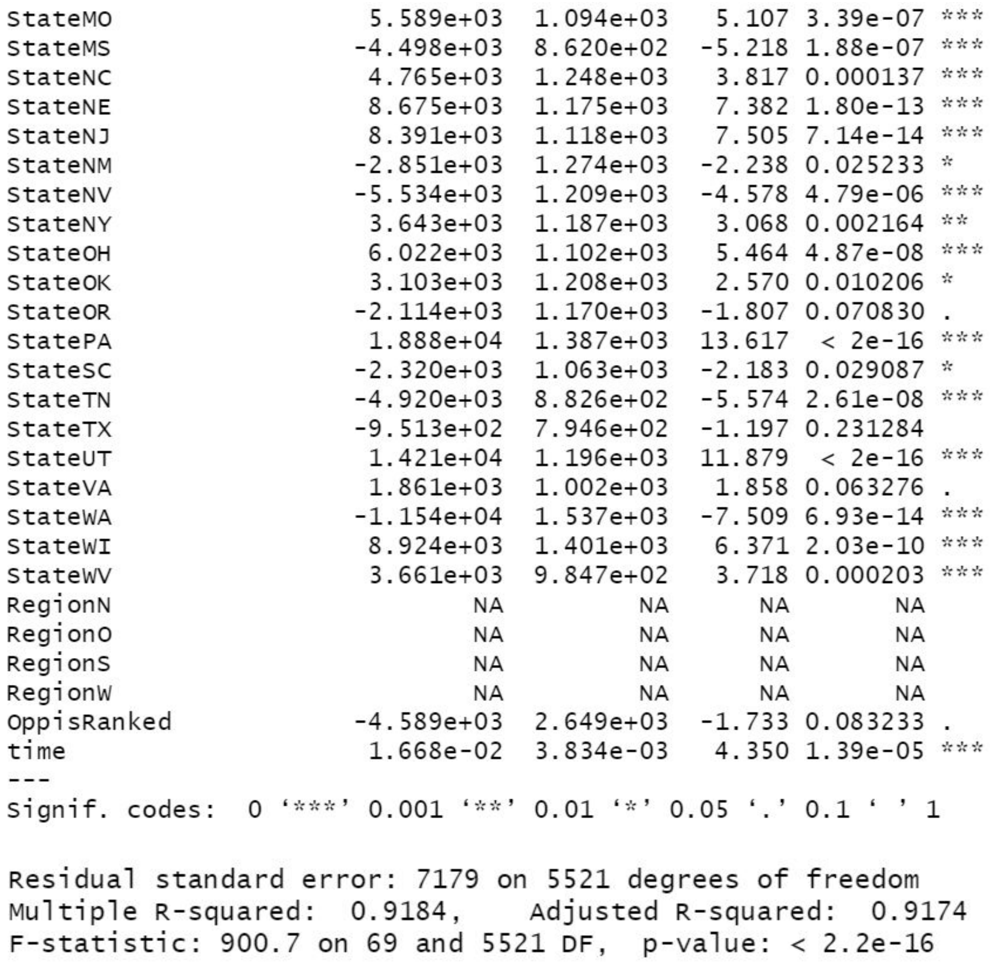
Figure 9: R Multiple Regression with All Predictors



```
Call:
lm(formula = Attendance ~ ., data = train)

Residuals:
   Min     1Q Median     3Q    Max 
-47124  -3998    436   4396  28810 

Coefficients: (4 not defined because of singularities)
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)            1.505e+05  4.914e+04   3.063 0.002202 ** 
Rank                  -5.264e+01  2.978e+01  -1.768 0.077188 .  
TV                     3.007e+03  2.768e+02  10.863  < 2e-16 ***
Current.Wins           5.048e+02  6.072e+01   8.313  < 2e-16 ***
Current.Losses        -1.032e+03  6.545e+01 -15.771  < 2e-16 ***
Stadium.Capacity       5.584e-01  1.295e-02  43.106  < 2e-16 ***
New.Coach             -4.675e+02  2.699e+02  -1.732 0.083389 .  
Tailgating             1.568e+04  6.635e+02  23.640  < 2e-16 ***
PRCP                  -8.557e+02  2.975e+02  -2.876 0.004037 ** 
SNOW                   1.026e+03  5.467e+02   1.877 0.060526 .  
SNWD                  -6.715e+02  4.081e+02  -1.645 0.099976 .  
TMAX                   5.867e+01  1.383e+01   4.241 2.26e-05 ***
TMIN                  -5.687e+01  1.493e+01  -3.808 0.000141 ***
Opponent_Rank         -8.704e+01  3.054e+01  -2.850 0.004382 ** 
ConferenceACC          1.359e+04  9.513e+02  14.287  < 2e-16 ***
ConferenceBig-10       2.975e+03  1.042e+03   2.853 0.004343 ** 
ConferenceBig-12       8.499e+03  8.959e+02   9.486  < 2e-16 ***
ConferenceBig East     4.440e+03  1.445e+03   3.073 0.002127 ** 
ConferenceCUSA         1.437e+03  8.152e+02   1.763 0.077916 .  
```



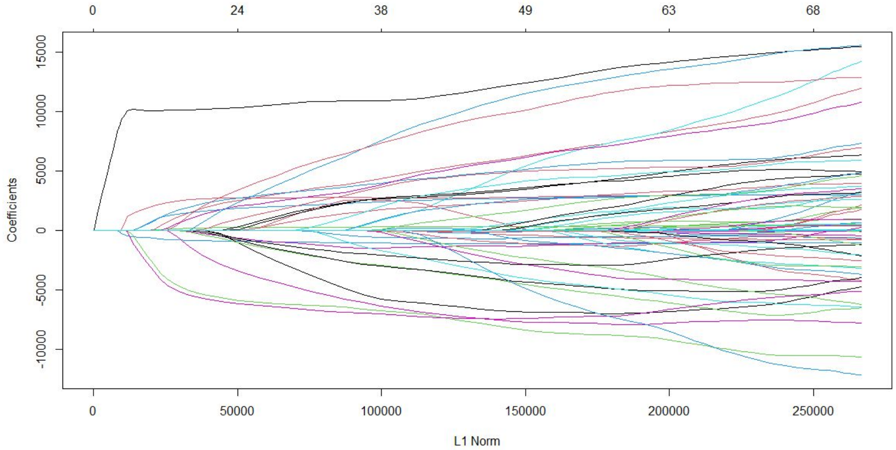
```
ConferenceFCS         -3.490e+03  1.212e+03  -2.880 0.003986 ** 
ConferenceIndependent  3.049e+03  1.194e+03   2.553 0.010714 *  
ConferenceMid-American -5.673e+03  9.949e+02  -5.702 1.24e-08 ***
ConferenceMWC          4.978e+03  1.031e+03   4.829 1.41e-06 ***
ConferencePac-12       1.624e+04  1.167e+03  13.922  < 2e-16 ***
ConferenceSEC          1.218e+04  1.024e+03  11.901  < 2e-16 ***
ConferenceSun Belt    -2.938e+03  8.735e+02  -3.364 0.000775 ***
ConferenceWAC          3.584e+03  9.183e+02   3.903 9.60e-05 ***
conference_game        4.562e+02  2.624e+02   1.739 0.082123 .  
isRanked              -1.304e+03  2.508e+03  -0.520 0.603062    
Special                2.969e+03  2.933e+02  10.123  < 2e-16 ***
GD                    -3.516e+03  1.435e+03  -2.450 0.014331 *  
Year                  -6.898e+01  2.434e+01  -2.834 0.004619 ** 
StateAR                6.469e+03  8.485e+02   7.625 2.86e-14 ***
StateAZ               -3.612e+03  1.291e+03  -2.799 0.005143 ** 
StateCA               -9.608e+03  9.674e+02  -9.932  < 2e-16 ***
StateCO                2.064e+03  1.143e+03   1.806 0.071001 .  
StateFL                3.797e+03  8.182e+02   4.641 3.54e-06 ***
StateGA               -7.923e+03  9.950e+02  -7.963 2.02e-15 ***
StateHI                1.146e+02  1.116e+03   0.103 0.918210    
StateIA                6.776e+03  1.153e+03   5.876 4.46e-09 ***
StateID                8.277e+02  1.162e+03   0.712 0.476277    
StateIL                2.973e+03  1.091e+03   2.724 0.006466 ** 
StateIN                3.360e+03  1.067e+03   3.149 0.001645 ** 
StateKS                3.163e+03  1.033e+03   3.062 0.002211 ** 
StateKY                4.814e+03  1.380e+03   3.488 0.000491 ***
StateLA               -9.252e+02  9.403e+02  -0.984 0.325161    
StateMA               -3.978e+03  1.016e+03  -3.917 9.07e-05 ***
StateMI                6.603e+03  1.409e+03   4.687 2.85e-06 ***
```



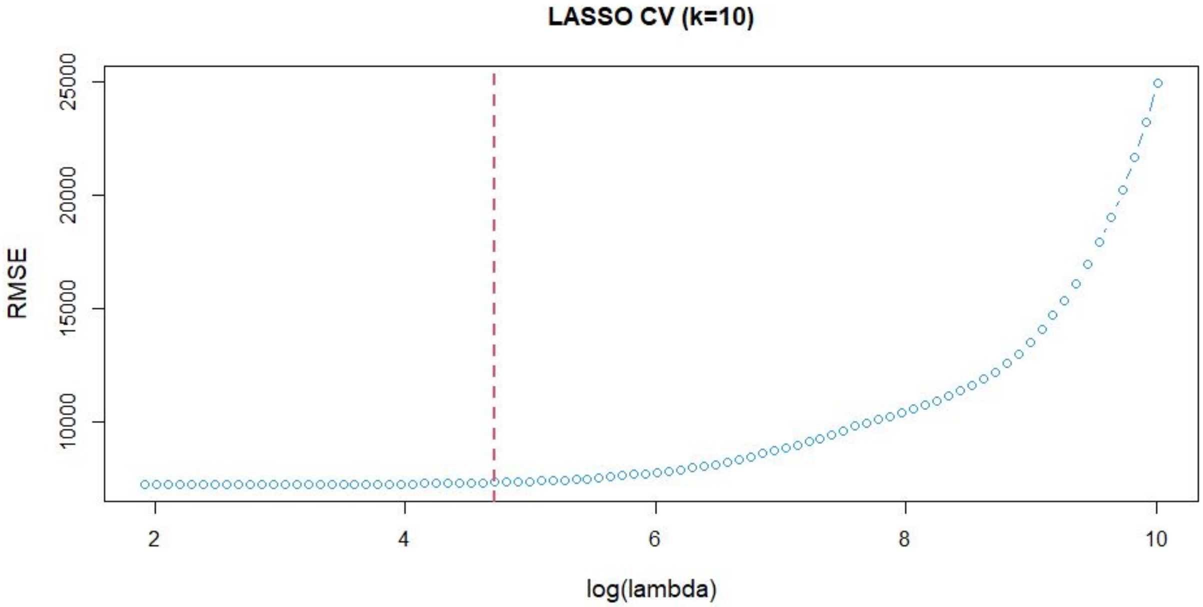
```
StateMO                 5.589e+03  1.094e+03   5.107 3.39e-07 ***
StateMS                -4.498e+03  8.620e+02  -5.218 1.88e-07 ***
StateNC                 4.765e+03  1.248e+03   3.817 0.000137 ***
StateNE                 8.675e+03  1.175e+03   7.382 1.80e-13 ***
StateNJ                 8.391e+03  1.118e+03   7.505 7.14e-14 ***
StateNM                -2.851e+03  1.274e+03  -2.238 0.025233 *
StateNV                -5.534e+03  1.209e+03  -4.578 4.79e-06 ***
StateNY                 3.643e+03  1.187e+03   3.068 0.002164 **
StateOH                 6.022e+03  1.102e+03   5.464 4.87e-08 ***
StateOK                 3.103e+03  1.208e+03   2.570 0.010206 *
StateOR                -2.114e+03  1.170e+03  -1.807 0.070830 .
StatePA                 1.888e+04  1.387e+03  13.617  < 2e-16 ***
StateSC                -2.320e+03  1.063e+03  -2.183 0.029087 *
StateTN                -4.920e+03  8.826e+02  -5.574 2.61e-08 ***
StateTX                -9.513e+02  7.946e+02  -1.197 0.231284
StateUT                 1.421e+04  1.196e+03  11.879  < 2e-16 ***
StateVA                 1.861e+03  1.002e+03   1.858 0.063276 .
StateWA                -1.154e+04  1.537e+03  -7.509 6.93e-14 ***
StateWI                 8.924e+03  1.401e+03   6.371 2.03e-10 ***
StateWV                 3.661e+03  9.847e+02   3.718 0.000203 ***
RegionN                        NA         NA      NA       NA
RegionO                        NA         NA      NA       NA
RegionS                        NA         NA      NA       NA
RegionW                        NA         NA      NA       NA
OppisRanked            -4.589e+03  2.649e+03  -1.733 0.083233 .
time                    1.668e-02  3.834e-03   4.350 1.39e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7179 on 5521 degrees of freedom
Multiple R-squared:  0.9184,    Adjusted R-squared:  0.9174
F-statistic: 900.7 on 69 and 5521 DF,  p-value: < 2.2e-16
```

Figure 10: Lasso Regression Model

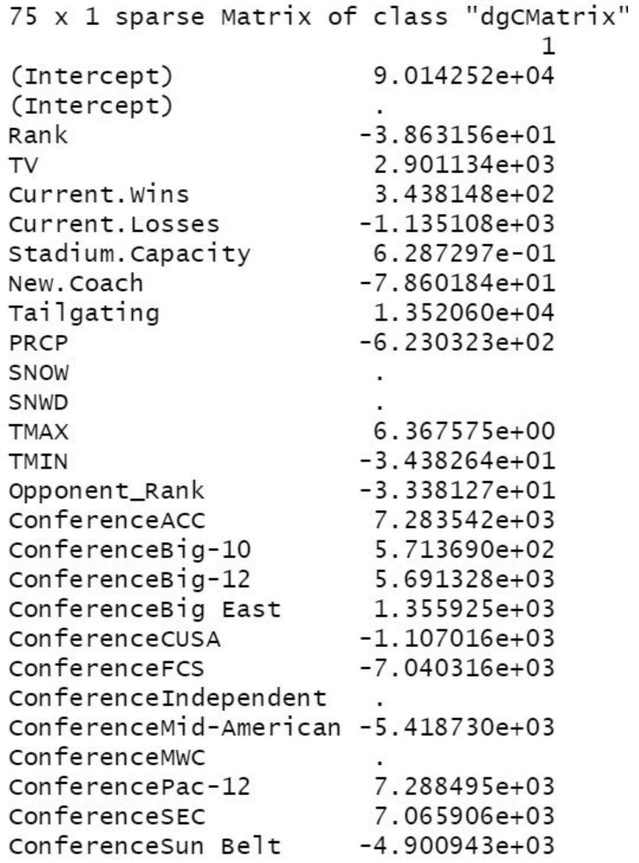
Lasso Regression Coefficients



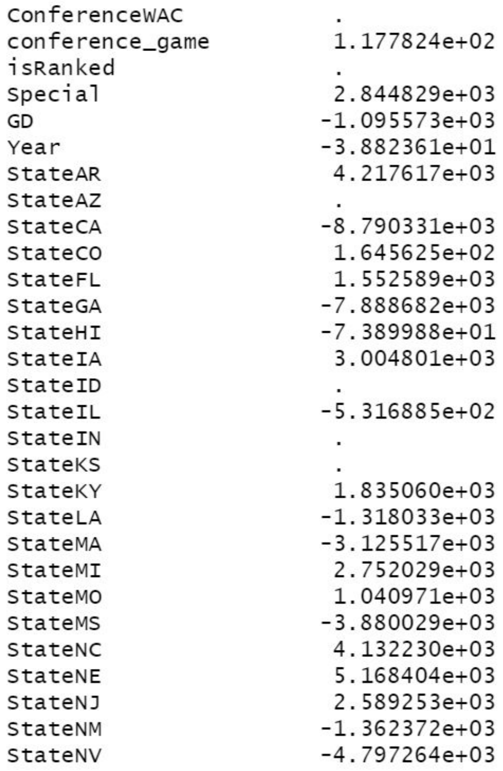
Lasso model with Lamba chosen by 1 standard error cross validation



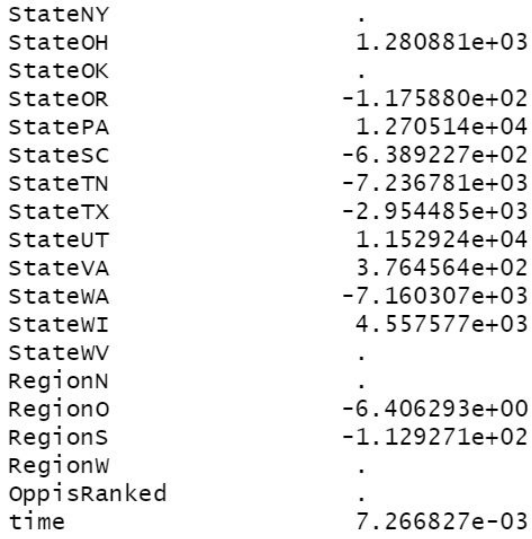
Lambda model coefficients



```
75 x 1 sparse Matrix of class "dgCMatrix"
                                       1
(Intercept)              9.014252e+04
(Intercept)              .
Rank                    -3.863156e+01
TV                       2.901134e+03
Current.Wins             3.438148e+02
Current.Losses          -1.135108e+03
Stadium.Capacity         6.287297e-01
New.Coach               -7.860184e+01
Tailgating               1.352060e+04
PRCP                    -6.230323e+02
SNOW                     .
SNWD                     .
TMAX                     6.367575e+00
TMIN                    -3.438264e+01
Opponent_Rank           -3.338127e+01
ConferenceACC            7.283542e+03
ConferenceBig-10         5.713690e+02
ConferenceBig-12         5.691328e+03
ConferenceBig East       1.355925e+03
ConferenceCUSA          -1.107016e+03
ConferenceFCS           -7.040316e+03
ConferenceIndependent    .
ConferenceMid-American  -5.418730e+03
ConferenceMWC            .
ConferencePac-12         7.288495e+03
ConferenceSEC            7.065906e+03
ConferenceSun Belt      -4.900943e+03
```



```
ConferenceWAC             .
conference_game           1.177824e+02
isRanked                  .
Special                   2.844829e+03
GD                       -1.095573e+03
Year                     -3.882361e+01
StateAR                   4.217617e+03
StateAZ                   .
StateCA                  -8.790331e+03
StateCO                   1.645625e+02
StateFL                   1.552589e+03
StateGA                  -7.888682e+03
StateHI                  -7.389988e+01
StateIA                   3.004801e+03
StateID                   .
StateIL                  -5.316885e+02
StateIN                   .
StateKS                   .
StateKY                   1.835060e+03
StateLA                  -1.318033e+03
StateMA                  -3.125517e+03
StateMI                   2.752029e+03
StateMO                   1.040971e+03
StateMS                  -3.880029e+03
StateNC                   4.132230e+03
StateNE                   5.168404e+03
StateNJ                   2.589253e+03
StateNM                  -1.362372e+03
StateNV                  -4.797264e+03
```



```
StateNY                        .
StateOH                        1.280881e+03
StateOK                        .
StateOR                       -1.175880e+02
StatePA                        1.270514e+04
StateSC                       -6.389227e+02
StateTN                       -7.236781e+03
StateTX                       -2.954485e+03
StateUT                        1.152924e+04
StateVA                        3.764564e+02
StateWA                       -7.160307e+03
StateWI                        4.557577e+03
StateWV                        .
RegionN                        .
RegionO                       -6.406293e+00
RegionS                       -1.129271e+02
RegionW                        .
OppisRanked                    .
time                           7.266827e-03
```

Figure 11: Range of attendance varies significantly at certain schools

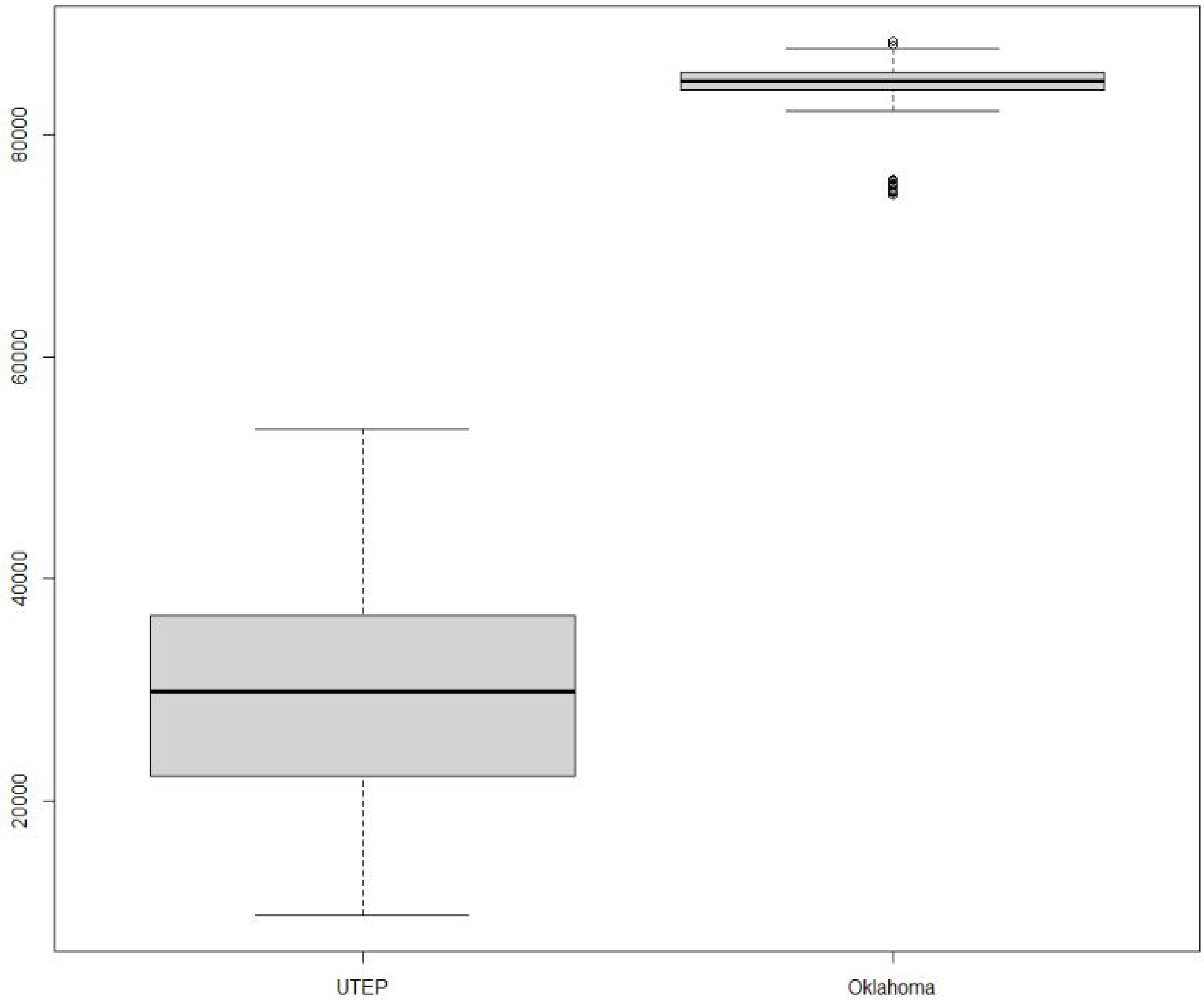


Figure 12: Texas A&M does show higher fill rates even after increasing stadium capacity and moving to the SEC in 2012.

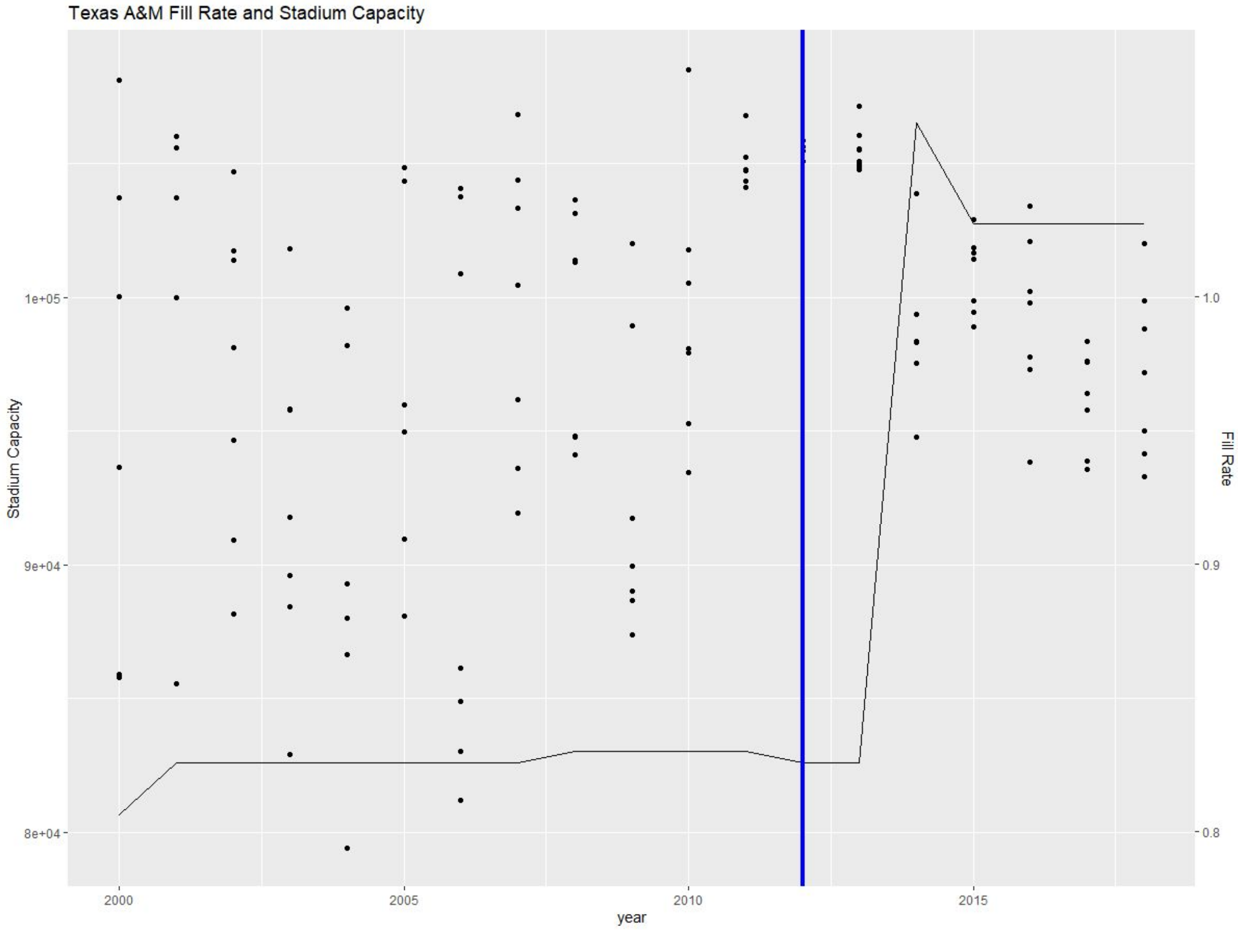


Figure 12: Range of Average Attendance for All Conferences

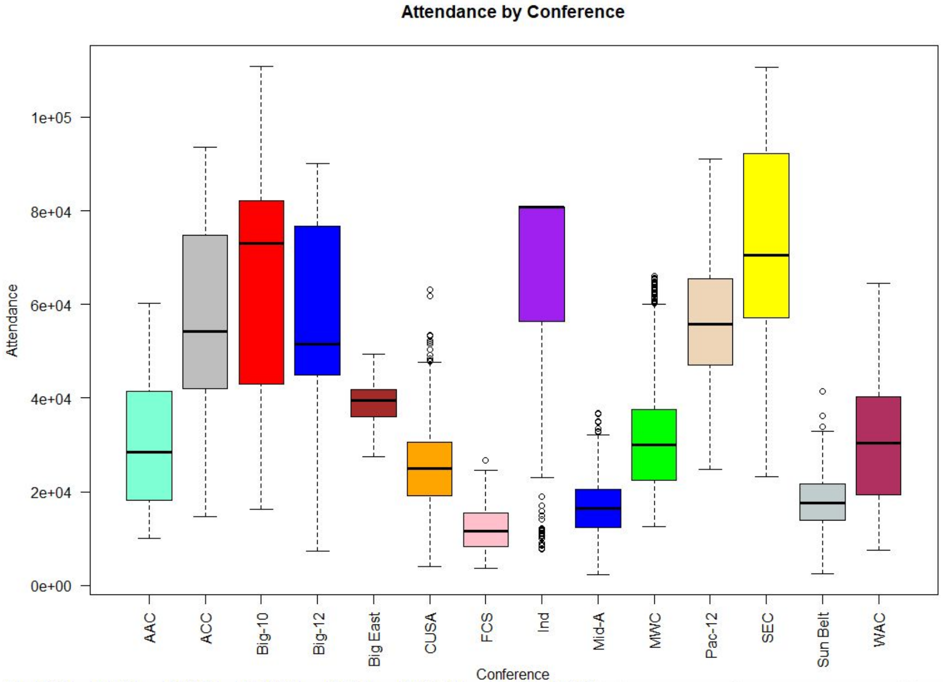


Figure 13. Range of Average Stadium Capacity by Conference.

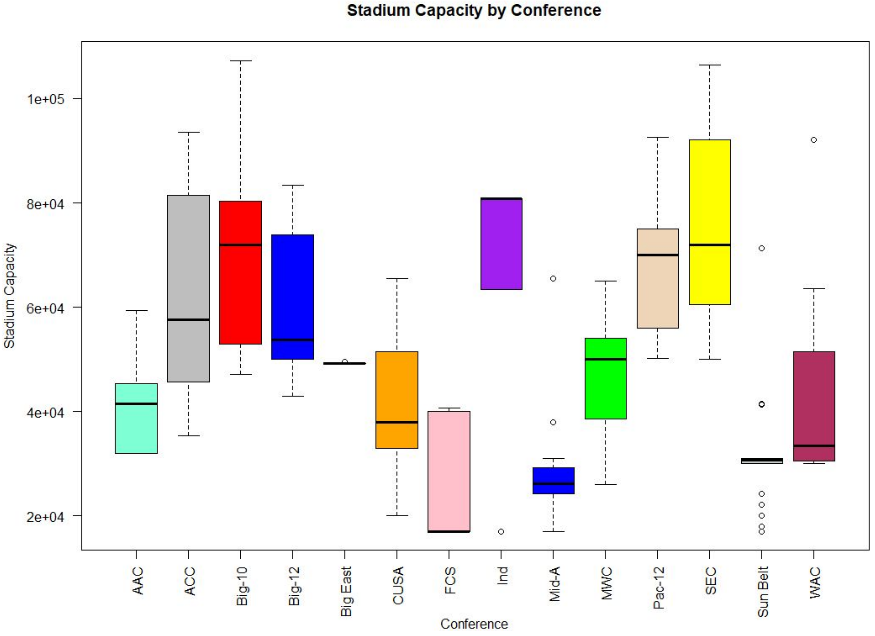


Figure 14. Range of Average Fill Rate by Conference

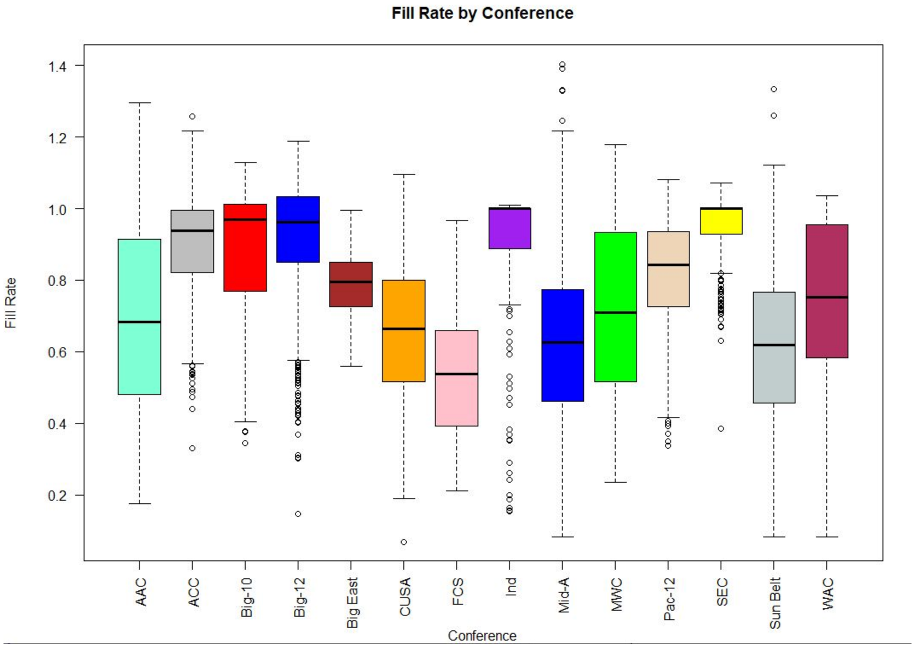


Table 1: College Football Data Columns

Each row represents a single college football game played between 2000 and 2018.
DateTime: Date and start time of game
Year: Calculated from DateTime, year game was played
Team: Home team playing the game
Opponent: Visiting team
Rank: Rank of home team, 1-25 if ranked in AP Poll, 99 if unranked
isRanked: Yes/No variable for whether home team was ranked in AP Poll
Opponent_Rank: Rank of visiting team, 1-25 if ranked in AP Poll, 99 if unranked
OppisRanked: Yes/No variable for whether opponent team was ranked in AP Poll
Site: Concatenated stadium name, city, and state
State: Calculated from Site, state game was played in
Region: Calculated from State, region game was played in (South, North, East, or West)
TV: TV network of game if on TV, converted to Yes/No dummy variable
Special: Yes/No variable for whether the match-up had any distinct features, for example, the Texas v Texas A&M Thanksgiving Rivalry game, or the Texas v Oklahoma Red River Showdown
GD: Yes/No variable for whether ESPN College GameDay was present on site
Attendance: Total in-person recorded fan count
Current Wins: Previous number of wins by home team
Current Losses: Previous number of losses by home team
Stadium Capacity: Listed stadium capacity
Fill Rate: Attendance divided by Stadium Capacity
New Coach: Yes/No variable for whether team had a new head coach
Tailgating: Yes/No variable for whether there was tailgating before game
PRCP: Amount of rain recorded during the game
SNOW: Amount of snow recorded during the game
SNWD: Yes/No variable for whether there was any snow during the game
TMAX: Maximum temperature in degrees Fahrenheit during game
TMIN: Minimum temperature in degrees Fahrenheit during game
Conference: NCAA conference of home team
Conference_game: Yes/No variable for whether the home and visiting team are in the same conference
Win: Yes/No variable for whether the home team won the game
Team_Score: Final score of home team
Opp_Score: Final score of visiting team
OT: Number of overtimes played