Joshua Harkness

STA6704

Project Write-Up

Professor Aaron Smith

**Overview**

The following analysis was preformed in an attempt to discover the strengths and weaknesses of the 2019 Jacksonville Jaguars. The goal through the analysis was attributes like the pass location during third downs could be used determine where the team’s strength and weaknesses are. It could also be used to determine trends. Trends do not automatically correlate to a strength or weakness, but if another team finds the trends first, it could quickly become a weakness. These insights could be used to provide the coaches with insights on where to draft or where the coaching organization may need some turnaround.

**Data Collection**

The initial data consisted of all NFL plays throughout the entire 2019 regular season from <https://github.com/ryurko/nflscrapR-data>. The original data consisted of 156 attributes and over 45000 observations.

**Data Preprocessing**

Firstly, the number of attributes were reduced based on personnel expertise in NFL football. Other attributes that I believed to be to specific for this analysis were removed afterward. The next step was to extract only the plays where Jacksonville was involved and the play type was either a run or pass.

A close up of a map

Description automatically generated

*Figure 1 shows the PCA analysis for the passing offense*

Afterwards, Primary Component Analysis (PCA) was run as another dimension reduction step. Through the PCA analysis, it was determined many of the remaining attributes were different ways of evaluating a similar dimension. The clearest example of this occurred when the quarter variable was removed in favor of keeping the game seconds remaining. Both attributes represented the time remaining in the game while the game seconds remaining allowed for more flexibility when the data was discretized. Some similar relationships were still left since it was believed they were still unique enough to help with the analysis. This in demonstrated in figure one, where although the downs are obviously not related, they could still provide insight further in the analysis.

The data was then manipulated so it could be directly related to Jacksonville. Once all the desired attributes were altered to reflect Jacksonville, the first dataset split occurred. The data was split into two datasets; one dataset for offense and the other for defense. The datasets were then cleansed to remove columns that were completely null. Attributes like pass location were not sensible for running plays. After removing null columns, some null values persisted.

The null values for the run data were the run gap when running up the middle. I corrected this gap to the center. There were other runs that could be up the middle but were marked as behind the guard, so logically, the only remaining position to run behind on the line was the center. For pass plays, sacks were removed from the analysis because it does not help determine where the pass offense it strong or weak unless we know where the defending playing broke through the line.

The next goal was to manually create a y variable to track the quality of a play. The quality of the play was based on an algorithm that considered the yards gained, location of the ball, first down gained, and turnovers. A mixture of each of these variables was considered to determine the quality of the play. For offense, the best plays were of quality zero while on defense the best quality plays were quality four. An example of this would be with turnovers, both the offense and defense obtained a quality rating of four. The offense should never give up turnovers and the defense should cause as many as possible.

Finally, the last data partition step occurred. The four datasets consisted of the pass offense, pass defense, run offense, and run defense. The normalization of the data occurred before each modeling phase because each model had its own requirements for when the data needed to be normalized.

**Data Modeling**

**Bayesian Models**

**A close up of a logo

Description automatically generated**

*Figure 2 shows the final Bayesian model for the passing offense.*

Bayesian models were used first as an attempt to get an immediate broad picture of the analysis. Blacklists were created to ensure obvious relationships, like “Quality\_0” and “Quality\_4”, would not be linked together since that did not provide any real insight. Many arcs through the analyses also had to flipped since the qualities could never lead to another attribute. Instead, the attributes lead to the quality of play.

Before visualizing each graph, the data was discretized and compared through constraint, score, hybrid, and local discovery based algorithms. Each dataset performed best when using an interval based discretization through a constraint based algorithm. The scoring was based on AIC to reduce the penalization on complex models.

While the graphs did show some meaningful relationships, they were often hard to interpret based solely on a visual context. When looking at the relationships behind the graphs, it was difficult to ascertain the full picture. Nodes were only related to their linked counterparts, so the only way to interpret the relationship was by stepping through multiple levels.

The amount of effort involved in reading each relationship did not appear worth it when looking at their ability to predict the quality of play. The Bayesian models were good at predicting plays at the extremes (Quality\_0 and Quality\_4) but suffered at predicting mid quality plays. When predicting a quality from one to three, all the predictions outputted they were not of the quality being predicted. The accuracy would still lead to between 70 and 80 percent though, because the number of plays in each quality was well split between the five qualities.

At this point, Bayesian models did not appear to help at predicting the quality of the play. The possible reason may have been an issue with the y variable I created. Instead of continuing using the quality attribute created, it appeared an unsupervised model may be the better option.

**K-Means Clustering**

Before each cluster analysis was performed, the optimal k value was determined based on the gap statistic, elbow method, or average silhouette width. Figure three shows the different analyses used for

A close up of a map

Description automatically generatedA close up of a map

Description automatically generated

A close up of a map

Description automatically generatedA screenshot of a cell phone

Description automatically generated

*Figure 3 Top left: Gap statistic over first 10 clusters. Bottom Left: Gap statistic over 40 clusters. Top Right: Average silhouette width. Bottom Right: Elbow method*

the passing offense. In most cases, the number of optimal clusters was four. The only analysis where the silhouette value was not clear was the passing defense. For the passing defense, five clusters were chosen since it was the initial number of different qualities.

The different clusters were then visualized to determine how well distinguished each cluster was. The passing offense and rushing defense had the most distinguished clusters, which can be seen in figure four.

A screenshot of a video game

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA picture containing traffic, air

Description automatically generated

*Figure 4: Top Left (a): Pass Offense. Top Right (b): Pass Defense.*

*Bottom Left (c): Run Offense. Bottom Right (d): Rush Defense*

Each dataset performed two k-means, but each time the lower number of clusters provided more distinguished clusters. The pass defense, depicted in figure 4b, was the most non-distinguished cluster. Looking at the clusters alone does not depict the actual relationship between the clusters though. To determine if the clustering visualized for the passing defense could be used, silhouette plots were used on each cluster. While this analysis was performed on each cluster, it was the most crucial for the passing defense based on the visualization.

Figure five illustrates that although the clusters appear to be bunched, the data was decently confident in the groups they belonged too. There are a few far negative values in each cluster, but most of the clusters consist of positive silhouette widths. Some of the silhouettes still had low values though, so the outcomes determined for the pass defense based on clusters one, four and especially five may not be too accurate.

A screenshot of a cell phone

Description automatically generated

*Figure 5 The silhouette widths for each cluster in the passing defense*

**K-Means Results: Passing Offense**

A close up of a newspaper

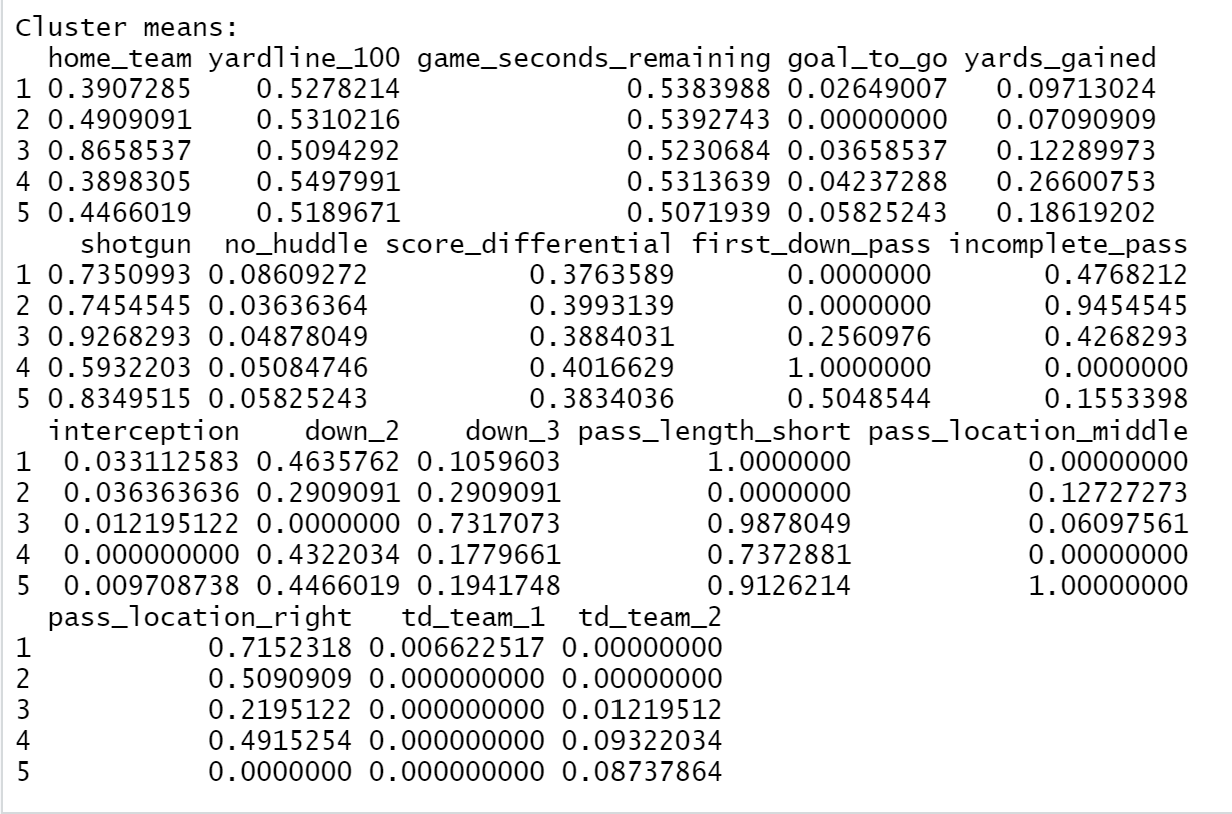
Description automatically generated

*Figure 6 Pass offense cluster breakdown*

Cluster four possessed all the first down passes and no incomplete passes or interceptions. The yards gained appears higher than all the other clusters and it is the only cluster with offensive touchdowns. Cluster four is clearly the most successful cluster for the offense but required further analysis to understand what made the offense so successful. There appeared to be more deep passes in this cluster than any others, but that was the only significant difference between this cluster and the others.

Another interesting relationship discovered in this dataset was between home and away games (clusters one and two). The Jaguars appeared to have far more first downs when playing away than at home. We can see this because all the downs in cluster one were closer to zero while most downs in cluster two were second downs. This appears to conflict with the idea of home field advantage. Also, when away, the team appeared throw to the right far more than when it was home and to the left when away.

Overall, the cluster analysis showed more positives overall for the offense than weaknesses, but it also displayed some tendencies that opponents may take advantage of.

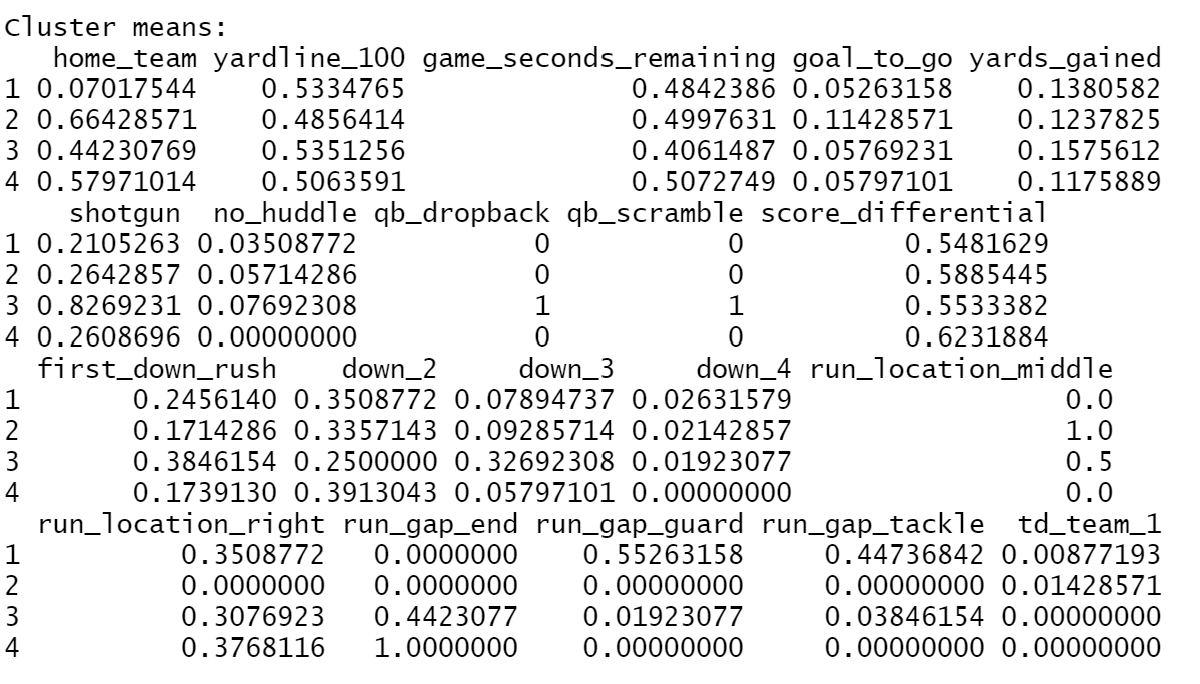


*Figure 7 Pass defense cluster breakdown*

When looking at defense, the values associated with first down passes and “td\_team\_2” should be low to be considered a successful play. Clusters four and five grouped together plays where the defense did not perform well based on the touchdowns and first downs given up. The reason for the poor performance is displayed in figure seven.

As discussed above, cluster five had the lowest silhouette scores so the analysis should be taken with a grain a salt. In cluster five, there was a strong relationship between short passes and the pass location in the middle. Cluster four did not appear to have any uniqueness to it that would explain a potential weakness for the defense other than it was the cluster with the least amount of shotgun plays. One explanation for this may be the defense was not as prepared for passing plays when the offense was not in an obvious passing formation.

Cluster three showed teams liked to pass out of shotgun more when playing against the Jaguars in Jacksonville and slightly preferred passing to the short left more than any other cluster when on third down. Cluster one shows short passes to the right were very unsuccessful, and the rest of the clusters appear to back this up too.



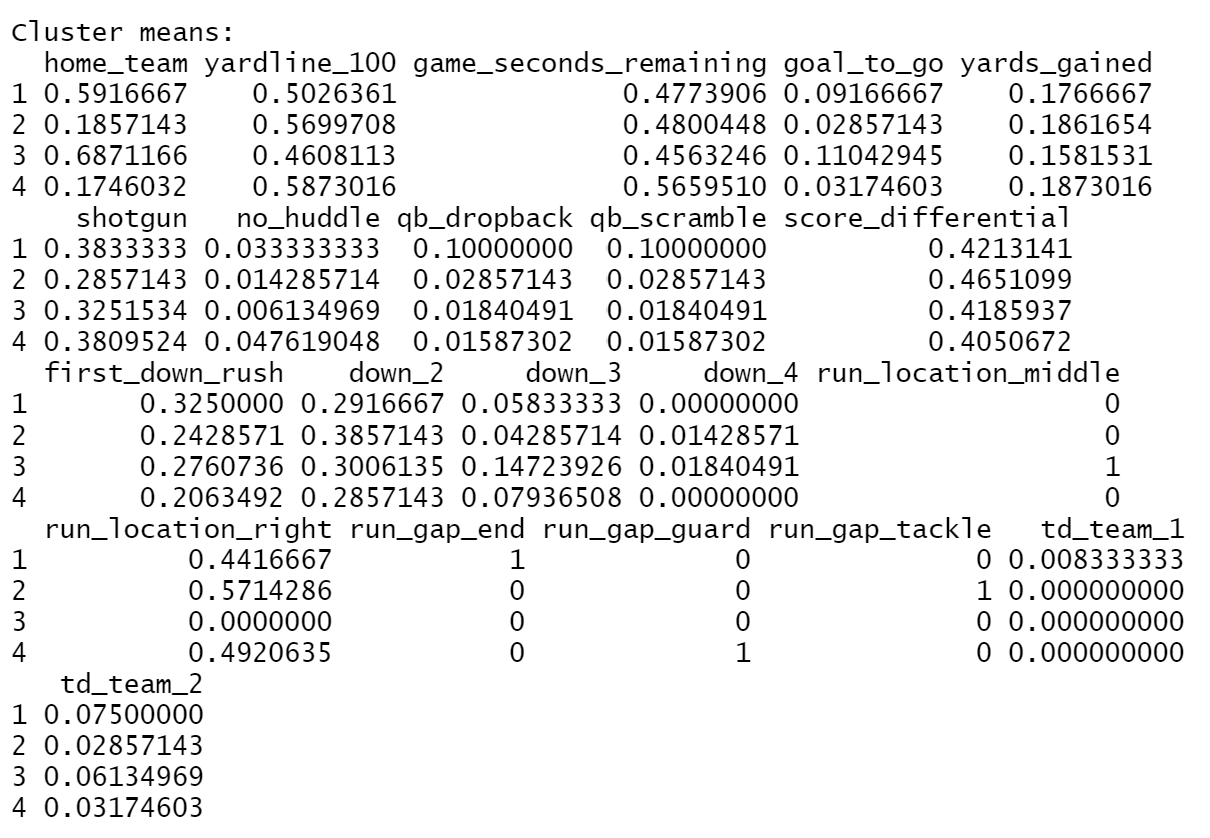
*Figure 8 Run offense cluster breakdown*

Figure eight shows there was not too much variability between the different clusters other than scrambles and away games.

Cluster three showed when the quarterback scrambled, which appeared to mostly happen on third down, often failed to lead to a first down although it was more successful at creating a first down the other types of rushes. This shows that the offense could have been putting the quarterback at too much risk and could possibly benefit from better receivers that get open faster or a better offensive line to allow the quarterback more time in the pocket.

Cluster one shows that the team is more likely to get first down rushes playing away. Many of the clusters throughout the different datasets appeared to show more success for the Jaguars when playing away. Cluster one and four looked similar yet have different first down results until the run gap was analyzed. While both clusters focused on runs to the left, it appeared the Jaguars were slightly more effective when running to the guard or tackle gap than when running around the edge. This may be an indication of either poor run blocking by tight ends or tackles. This may indicate the interior of the line was better at run blocking than the outside line.

Cluster two produced the most touchdowns with runs up the middle, but it appeared these runs were much more likely called when the team was already near the endzone (goal\_to\_go). This does reinforce the idea that the interior line was stronger at run blocking. When the team needed a tough run, they ran up the middle rather than bouncing to the outside.



*Figure 9 Run defense cluster breakdown*

All the clusters for the run defense analysis appeared to be based on the running gaps, as figure 9 shows. Plugging up run gaps in the NFL is key to stopping the run, so it was likely the clusters were showed how well the Jaguars did at defending each gap.

Cluster one showed runs to edge. The most first downs and touchdowns were given up when opposing teams ran to the edge. Slightly more runs to edge were also to the left side of the defense, indicating the left outside linebacker or defensive lineman may not have been fast enough to seal the edge.

Cluster two showed when teams do run to the tackle gap against the Jaguars, they tended to run more to the right, likely because the outside defensive lineman was weaker there.

Cluster three represented runs up the middle, which was where the second most touchdowns and first downs against the team occurred. Runs up the middle also appeared to occur more earlier in the game than later but had the least yards gained. This showed that teams were more likely to run the ball up the middle in short yardage situations and are decently successful at it relative to the other runs.

Cluster four contained the guard gaps. These gaps appeared to be the most solid for the defense on both the left and right side. Although the Jaguars gave up the most yards (slightly) to the guard gaps, they gave up the least first downs and second least touchdowns. The combination of cluster two and cluster three was a bit contradicting because they both focus on the interior of the defense. This may be because the defense had strong linebackers in the middle to plug up the gaps, but maybe weaker interior linemen to prevent successful runs right up the middle.

**Condensed Results**

|  |  |  |  |
| --- | --- | --- | --- |
| Passing Offense | Passing Defense | Rushing Offense | Rushing Defense |
| * More success occurred when throwing the ball deep. * More first downs when playing away. * Throw to the right much more when playing away. | * + Weak against passes to the short middle   + Weak against passes not from shotgun   + Weak against short left passes on third down, especially when home   + Very strong against short passes to the right   + Strong against deep outside passes | * + Scramble too much on third downs without great success   + More first downs when running in away games   + Weak on the edges   + Strong at the tackle and guard gap | * + Weak edge rush, especially to the left   + Weak right tackle gap   + Good at preventing first downs in the guard gaps. |

**Conclusion**

For this study, the cluster analysis seemed more informative. The creation of a unique y variable likely should not have been done, and an unsupervised model should have been immediately used.

The analysis shows a few tendencies and weak spots the Jaguars may be able to focus on during the draft or free agency. The run defense did not show many strengths, so that may be the first place to look. The team also ranked near the bottom of the league in rush defense for the 2019 season, which further backs up the analysis of the run defense.

Outside of the run defense, it looks like the team had a few obvious weaknesses, but possessed some good strengths too. After these weaknesses are addressed, the team may be strong enough to make a playoff run.