Project Title: Uncovering Trends Behind Success in the NFL

Team Name: Scott's Tots

Leader: Curtis Hsu

#### **Team members**

Member 1: Curtis Hsu (<u>curtishsu@college.harvard.edu</u>) Member 2: Ted Zhu (tedzhu@college.harvard.edu)

Member 3: Ray An (dongzhengyangan@hsph.harvard.edu)

Member 4: Kelly Mcdougall (kmcdouga@mit.edu)

10/02/18 - Decide on a data set; decide on final project topic

10/07/18 - Think of ideas for audience, specific goals, questions, and story to tell

10/09/18 - Create sketches and storyboard; complete detailed project plan (Page 4)

10/16/18 - Met with Michael to discuss story for project

10/18/18 - Created prototype v1

10/23/18 - Created prototype v2

10/29/18 - Met with Michael and polished prototype v2 for final project

#### Week 1 - 10/02/18

#### **Project Abstract:**

(The abstract should include your motivation for doing this project, what you are trying to do and which goal you want to achieve. Also include a description about your planned datasets and where you will acquire them.)

The motivation of this project is to take a dive into three teams over the last 8 years and see what variables correlate with their success or failure. We are trying to see if there are any noticeable actions that certain teams conduct that allows them to consistently perform well. The goal is to clearly delineate what good teams do that bad teams fail to do. The planned data set is play-by-play in the NFL from 2010 - 2017, which captures the situation factors of each play in every game. It is scraped using the NFL scrapR package.

#### Week 2 - 10/07/18

#### **Analyze audience:**

The audience for this webpage will be general football fans who are interested in which variables differentiate stronger teams from weaker teams

#### **Specific Goals:**

- Use visualizations to compare three teams over the course of 8 years worth of data
- What makes a good team good and what makes a bad team bad?
- Explore the differences in teams along themes of red zone performance, field positioning, third down success, and personnel
- Do a longitudinal study of 7 years of data to determine what variable(s) correlate to significant changes in team performance over time.
- Specifically investigate 3 teams (Patriots, Browns, Lions) known to be consistently good, bad, and mediocre

#### List of (at least) 10 Questions our visualizations should be able to answer:

- Difference in run/pass distribution
- Difference in number of players represented at the quarterback position
- Difference in number of plays being run on offense
- Distribution of time on field for offense and defense
- Difference in Performance of passer and receiver in three teams
- The distribution of the total touchdown compared in three teams
- The network between the passer and receiver in the three teams
- The performance compared with HomeTeam and AwayTeam
- The touchdown probability, field goal probability in the three team
- The number of wide receivers thrown to and how many yards per receiver
- Red zone performance; from what starting points do teams score
- Scoring performance: on what yard line do what proportion of drives result in touchdowns? What about field goals?
- Third down conversion what is success based on distance to marker
- How often do the Patriots find themselves down in win percentage in the first quarter vs other teams?
- When the Patriots get the ball back, what is the expected points scored on when the drive first starts?

## Describe what kind of story you would like to tell. Come up with some candidates for your main messages.

From a story perspective, we want to do a team by team comparison, specifically looking at the New England Patriots, the Cleveland Browns, and the Detroit Lions, as models of sustained excellence, sustained bottom of the barrel performance, and mediocrity in the NFL.

First, we'll start off by introducing mostly narrative information, by focusing on the Patriots. Highlight their successes in the form of Super Bowl wins, postseason success, and records being broken. Then, we'll focus on the Browns and their terrible performances, mostly in the form of their 1-31 record over the last two years. For this, we'll make basic visualizations to highlight key figures over the last 5-10 years and link in videos from sources such as espn.com or nfl.com

From there, we'll start to use our play-by-play data to slice and understand what is really going on. We'll examine a few major themes that will help us differentiate. These themes are, for both offense and defense: field position, consistency of personnel, third down performance, red zone performance, and overall major stats (pass yards, total yards per game etc).

## **Detailed Project Plan**

## Names of those who actively worked

- Curtis Hsu
- Ted Zhu
- Kelly Mcdougall
- Ray An

#### **Goals and Tasks**

The goal of this project is to take a deep dive into three NFL teams and effectively conduct an 8 year longitudinal study on each team. We will take three teams which we believe to represent models of sustained excellence, sustained bottom-of-the-barrel performance, and mediocrity in the NFL (New England Patriots, Cleveland Browns, and Detroit Lions). On a high level, we hope to tell the story behind the success of each team. More specifically, we want to use visualizations to highlight correlations behind the success of the team and different variables such as run / pass split or player turnover. By doing this for each of these three teams, we will tell the story of what makes this team good or bad. As a result, the general audience for our project will be American football enthusiast who is interested in the numbers behind the success of these teams.

The first task we should accomplish is highlight the differences in success of each of these three teams. First, we'll start off by introducing mostly narrative information, by focusing on the Patriots. We will highlight their successes in the form of Super Bowl wins, postseason success, and records being broken. Then, we'll focus on the Browns and their terrible performances, mostly in the form of their 1-31 record over the last two years. For this, we'll make basic visualizations to highlight key figures over the last 5-10 years and link in videos from sources such as espn.com or nfl.com. Finally, we will highlight the mediocre performances of the Lions. To accompany these text narratives, we will use an effective visual to convey success through data, such as record per season over time.

Our next task is to identify and highlight key variables that we believe will influence success in the NFL. To tackle this task, we first identify few major themes that will help us differentiate. These themes are, for both offense and defense: field position, consistency of personnel, third down performance, red zone performance, and overall major stats (pass yards, total yards per game etc). To attain this information, we will use play-by-play data from 2010 to 2017. We will

cut the data based on the different themes highlighted above. Once we have identified the key variables that support each theme, we will structure our web page accordingly. For our next task, we will construct visuals according to the identified variables above. Using these themes as guides, we will create data-driven visuals for each theme. For each team, we will

use a combobox that allows the user to choose which team to analyze. We will do this under 4 different views: one for each team, as well as a combined view. In doing so, we will be able to conduct analysis under each team. In addition, we will attach a text of analysis to each visual.

Finally, we will include a paragraph that synthesizes the analysis from each visual. That is, we will holistically analyze which variables of each theme are consistent over time for each of the teams. In doing so, we can determine which underlying trends are driving success for these teams.

### **Description of the data**

**Obtaining the data:** The data has already been scraped using the NFLscrapR package in R. This package was created by Carnegie Mellon for play-by-play analysis in the NFL.

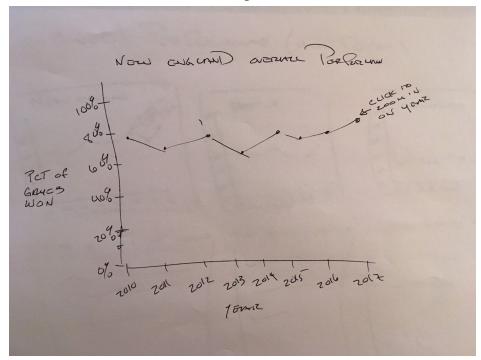
**Description of the data:** The data used for this project is play-by-play data in the NFL from the years 2010 to 2017. The the play-by-play data includes situation factors of the play, such as which teams are playing, which yard line a team is currently at, how many yards until first down. The data also includes play information, such as whether the ball was run or passed, to which direction, how many yards gained, and and active players involved. Finally the data also includes overarching ideas such as which teams are playing, which stadium, and when.

#### **Sketches**

Begin by providing the user with an overall view into the long term performance of each team (y: winning percentage, x: years ) and position them side-by-side. Below each table there could be some brief background information about the team ("New England has historically been good" "Cleveland has historically performed poorly...")



View of each table with the team's performance over time:

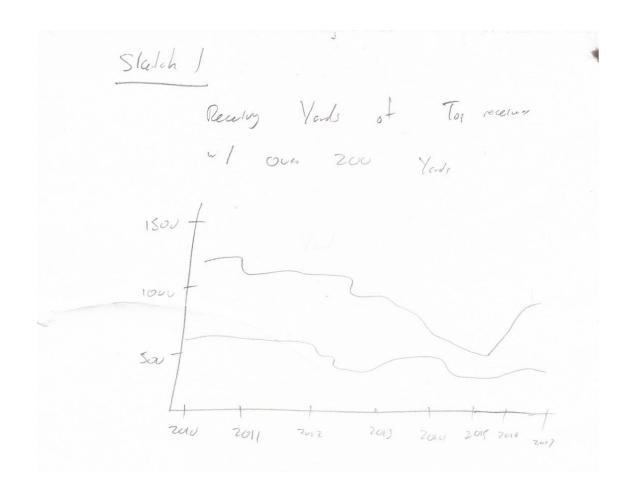


Within each team visualization the user could select a particular year to see each game that year

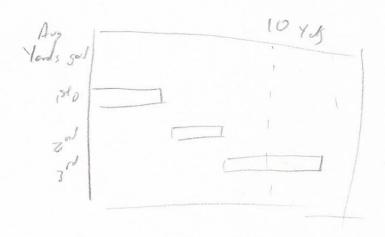
Also could make use of the brush component to allow users to select a specific date range, which would cause the related data to update accordingly (similar functionality as HW5):



Sketch 1 below highlights the receiver diversity of each team. Sketch 2 highlights the performance of each team on 1st down, 2nd down, and 3rd, and benchmarks this against the 10 yards needed for a first down conversion. Sketch 3 shows the red zone performance of teams.



Statel Z



Sleetch 3

Red Zone Parfun

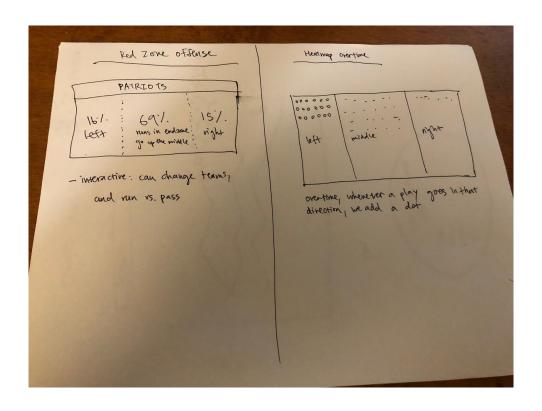
TO warn Nel

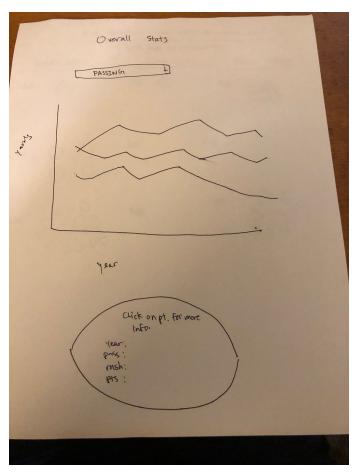
(00% Gree stat

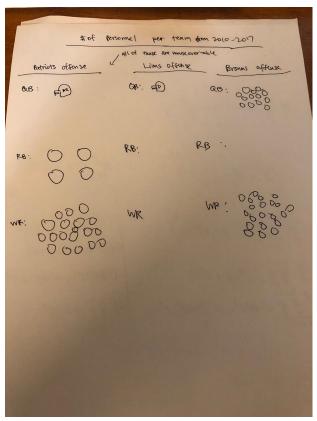
80%

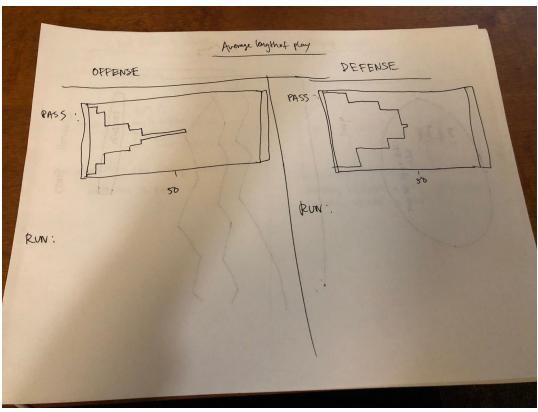
50%

30%









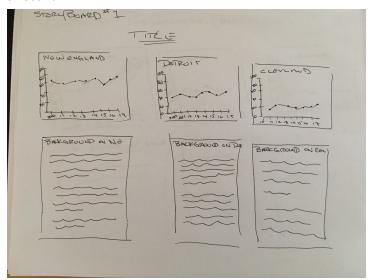
**Interaction Storyboard** 

#### - Sketch/View #1:

- Overview: Lands the user in a top-level view of the overall, long-term performance of each team. Provides the user with the winning percentage for each team between 2010 and 2017. Each chart is interactive and provides a tooltip (year, winning percentage for that year) and each year is clickable. Clicking the year brings the user to a chart specific to that year. (Question: do we want to have each chart update to maintain the comparison or focus on that particular team?)

Below each chart there is also some narrative text about each team.

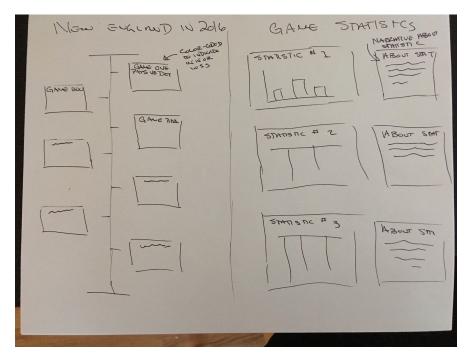
#### - Sketch:



#### - Sketch/View #2:

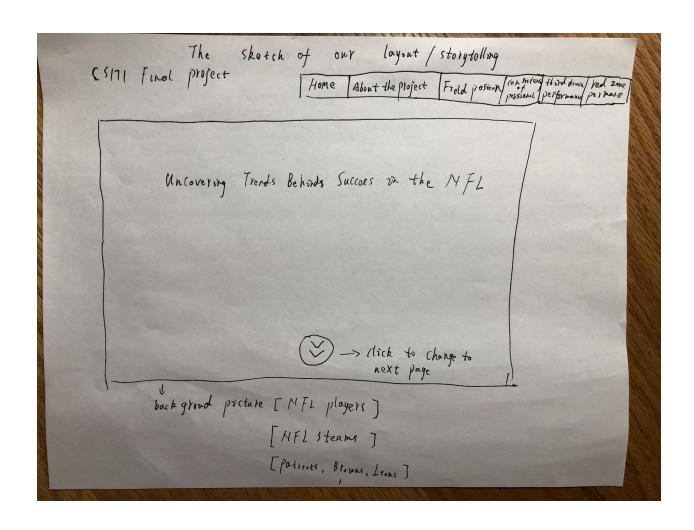
Overview: Once inside the specific year, the chart becomes a linear timeline displaying each game. The timeline would include basic information for each game (i.e opponent, game number, date of game, final score) and could be color-coded to indicate win or loss. The visualizations on the right would provide more detailed information and be accompanied by brief explanatory text about why that particular statistic is important.

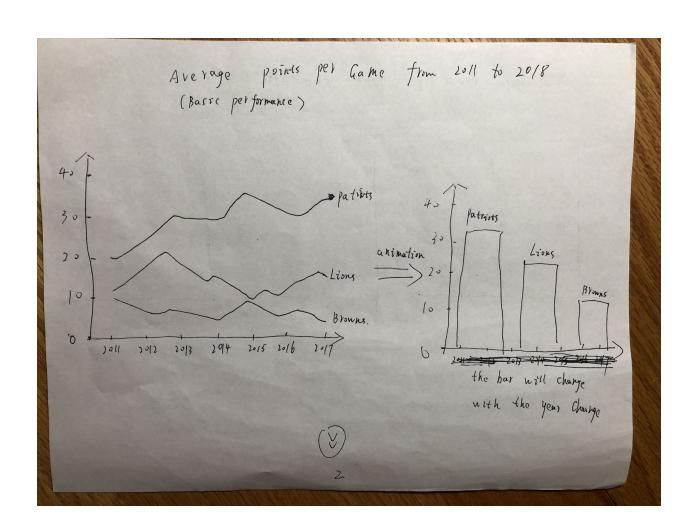
#### - Sketch:

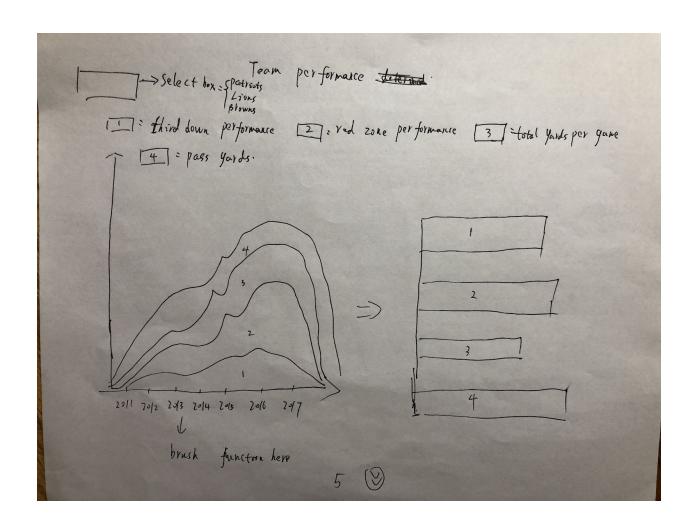


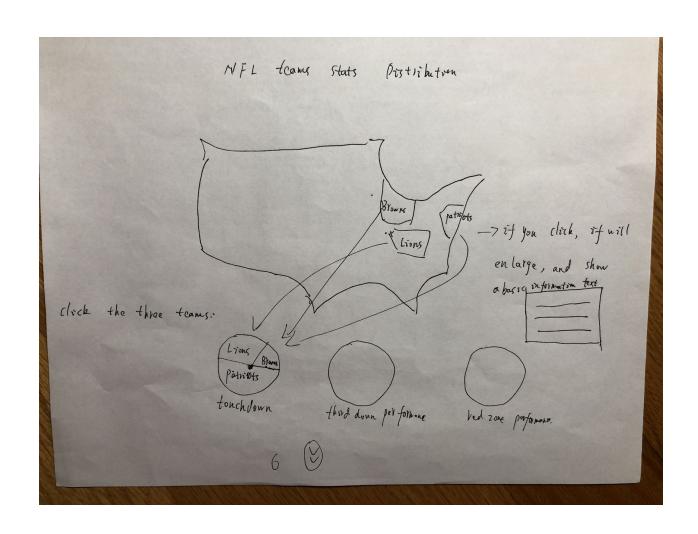
## Webpage layout/storytelling:

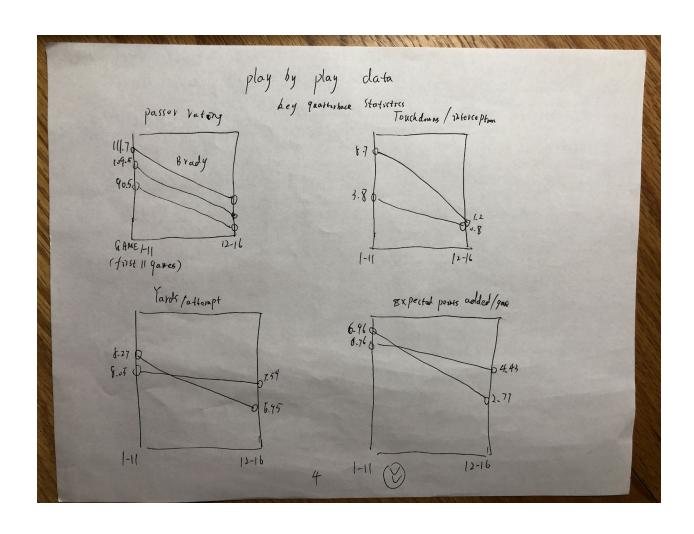
The story told is to start from the top and convey the team's overall performance. We will then move through subheadings to explore the factors behind performance. This will start with overall statistics of the team. We will then shift more granularly to the specific plays of each team, and then the players.

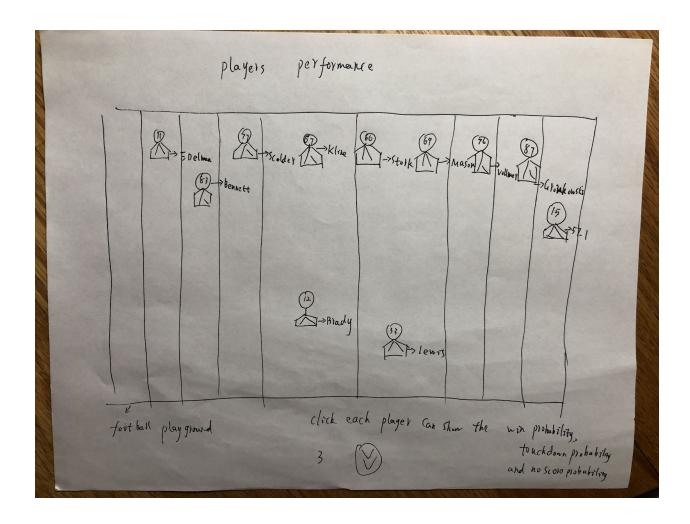












# A project timeline (with milestones when you are planning to finish which feature)

#### Prototype V1 due 11/18:

- Expectations: "For your prototype V1, we expect you to hand in your code and your process book in its current state (see Final Project Submission for details). You should have completed your data acquisition, and you must have a working visualization prototype. You do not have to have all your views up and running, and it does not need to be completely interactive, but the direction and the content must be clear."

#### - TODO:

- Rough website outline separated into the following portions:
  - 1. Intro including description of narrative portion and linking in videos
  - 2. Overall Major Stats (pass yards, total yards per game)
  - 3. Field position

- 4. Consistency of Personnel
- 5. Third down performance
- 6. Red zone performance
- Each of the following portions should have at least 2, preferably 3, pieces of media in it (videos, images, visualizations etc, with preference for visualizations but not necessarily strict). Portion 2 should have more visualizations.
- Visualizations do not need to be complex but should highlight a few key stats

#### Prototype V2 (due date is yet to be specified):

- Expectations: "For your prototype V2, we expect you to be 95% done with the implementation, and have your process book up to date. After prototype V2, you will only have one week to incorporate comments from peer feedback to produce a final interactive project website."

#### - **TODO**:

- Discussion about if we should add another visualization for each section
- Discussion for how we can improve existing visualizations to be more complex/creative
- Each person will take ownership of their category and "upgrade" or add visualizations
- Besides the visualizations, flesh out any narrative portion to help tie together the disparate parts of the website

## A feature list (with must-have, good-to-have, and optional items)

- Visualization(s) showing long term performance of each team. Possibly interactive allowing users to see detailed information for each year?
- <u>Features to slice by</u>
  - Drive
  - Quarter
  - Down
  - Ydstogo For examining red zone opportunities (exclude plays where this value is 0 because this represents a kickoff)
  - Posteam for which team is on offense
- Key features (core features we must be incorporating):
  - Yards.Gained yards gained on the play (sum up over the drive to see total)
  - Desc description of the play, need parsing to figure out if the field goal is good
  - PlayType important for checking what type of play is happening

- Passer, Rusher, Receiver (pair with use of Passer\_ID, Rusher\_ID, Received\_ID to make sure things are consistent)
- Touchdown (boolean)
- FieldGoalResult
- InterceptionThrown
- Features to consider adding in when we explore each theme
  - PassOutcome, PassLength, AirYards, Pass Location
  - RunLocation, RunGap
  - Fumble (boolean)
  - AbsScoreDiff calculated with respect to the home team, can be useful to see how much time/how many plays does a team spend in the lead or behind
  - Win Prob use same as above
  - WPA Win probability added (per play)

## A description of team roles

Overall, we will all actively participate in the active making of the website.

Curtis/Ted: Split the Intro, Major stats, and Red Zone performance

Ray/Kelly: Split Field position, Consistency of Personnel, Third down performance

<u>Process:</u> For each visualization you will create, write the specific question you will investigate below, the category it belongs to, and a brief sentence or two about what shape this visualization will be. That way, we can know what you are working on and provide feedback for each other both in terms of question asked and design visualization choices. Reference a sketch if it is attached in this doc.

#### **Example:**

Major Stats:

Question 1: What are the major stats between teams over the years?

-Use a multi-line chart to compare over time between teams. Dropdown menu will allow us change between stats (pass yards per game, rush yards per game, average score per game etc).

#### Intro

- Record of Patriots; record of Browns, picture of Lombardi Trophies / awards
- Three column describing the season of Pats, Lions, and Browns
- Include video
- Soccer visual
  - Line Graph: record of each team over 8 years
    - Click on a year
  - Filter for Bar graph of major stats
    - Pass yards per game
    - Rush yards per game
    - **■** Total points scored per game
    - Penalty yards per game

#### **Major Stats**

- Stacked area chart: How points are scored?
  - o Touchdowns, field goals, safeties, XP
- Stacked bar chart
  - Offensive plays, defense plays
    - Runs, passes, others

#### Third Down Conversion

Line chart to compare three teams' Third Down Conversion from 2010 to 2017 Animation visualization from Line chart to bar chart and pie chart, add brush function

#### Red Zone

- Histogram where each bucket is yard and want proportion of touchdown conversations
- Run direction / pass direction in red zone

#### Personnel

Compare the frequency of change of passer among these three teams Show the main passer's pass outcome (Tom Brady)

#### **Field Positioning**

\_

Yard to go

## 10/16/18 - Meeting with Michael to discuss direction of project

- Chatted with Michael <u>decided that the story of the webpage will be "How to build a playoff team"</u>
- The main visualization will be a decision tree
- And then each supplemental visualization will be based on something interesting we see from a node of the decision tree

#### <u>Technical Notes (Decision Tree building):</u>

Feature names (as in the dataset):

"PointsPerGame x", "ThirdDownPercentage", "RedZonePercentage",

"OffensiveYardsPerPlay", "PenaltyYards", "TimeOfPossession", "TurnoverDifferential",

"RushingYards x", "Times Sacked x", "PassingYards x", "OpponentRedZonePercentage",

"OpponentPassingYards x", "OpponentRushingYards x", "PasserRating"

Feature names (vernacular):

Passing yards per season

Rushing yards per season

Passing yards let up per season

Rushing yards let up per season

#### Third down conversion rate - got it

## Average yards to go for 3rd down - can be part of the "story" for third down conversion rate

Distribution of plays when it comes to direction

Average starting field position

Diversity of personnel for each play that happens under receiver, quarterback, and running back Head coaching change/offensive or defensive coordinator change

#### Average yards per play - got it

Average time spent on offense/defense

## **Prototype V1 Write-up**

#### Name:

- Curtis Hsu
- Ted Zhu
- Ray An
- Kelly Mcdougall

#### **Data Scraping and Cleaning**

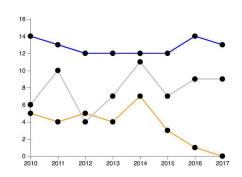
The data play-by-play data was scraped using the NFLScrapR package and from fantasydata.com. We obtained offensive and defense statistics, as well as playoff status, of each team from 2010 to 2017 and merged them using pandas. As the narrative of our project is to "how to build a playoff caliber team", part of the data mining process is building a decision tree to see which main features build a playoff caliber team. We have finished building that classifier, with the visual of the tree below under "First design of innovative view."

#### At least 1 visualization implemented + drafts for more

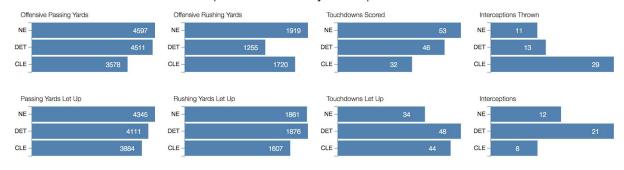
#### Visualization 1 (implemented):

-This visualization will be replaced as the main visualization with a d3 implementation of the decision tree shown below.

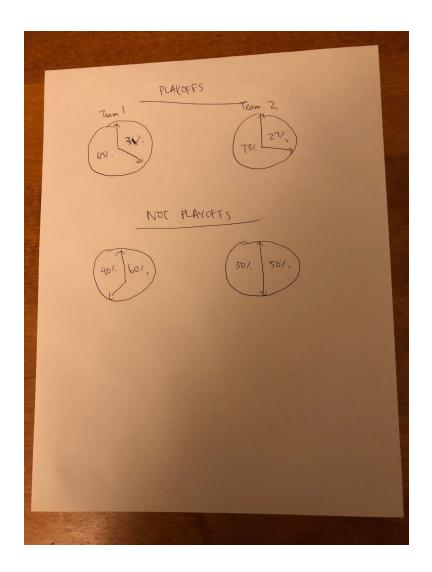
## Main Visualization



## Click data point to see year specific details

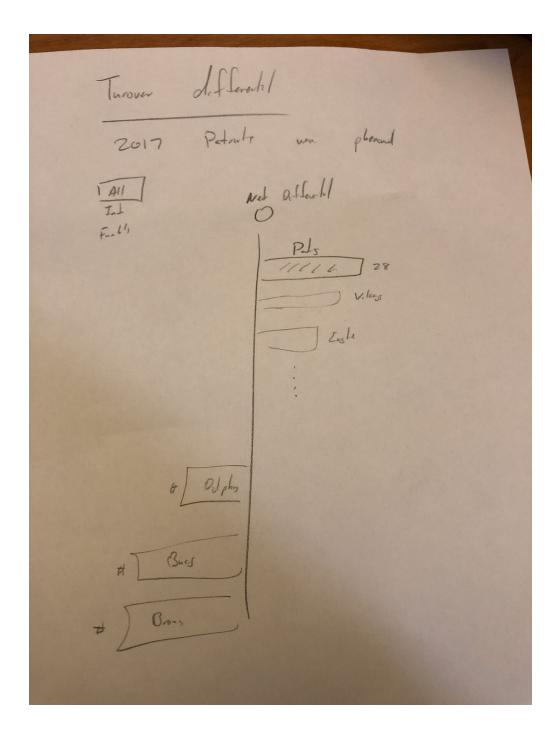


## Visualization 2:



The clock visual is supposed to match the theme of the time, where we are exploring the distribution of the time of possession. Since all teams play the same total of time, the circles are perfect since everything is standardized to the same total. We will compare top 5 and bottom 5 teams in terms of this stat for a particular year to show how this stat affects making the playoffs.

#### Visualization 3:



To encode turnover differential, we will highlight a team who performed extremely well on turnover differential and juxtapose it with other high performing teams and a few low performing teams. Because turnover differential can be both positive and negative, we will use a horizontal spanning bar chart with strong teams with positive differentials at the top and those with negative at the bottom. We will also include a dotted line to show where the threshold from the decision tree lies.

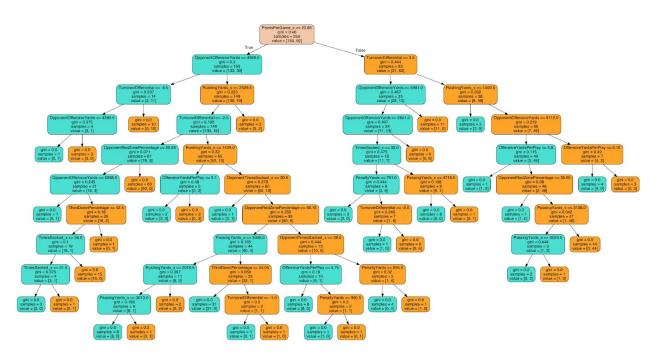
#### Rough webpage design and structure:

Webpage design and structure implementation is completed and can be accessed through index.html in the "MajorStats2" file of the submitted implementation folder. The portions in the white boxes are placeholders for information/visualizations based on interesting things we are choosing from the final implementation of the decision tree.

#### Story-telling:

The main story of our webpage is "how to build a playoff caliber team," as discussed during our meeting. To start, we will build a decision tree, in which the input are various statistics about a teams performance and the output is whether or not they made the playoffs. From building this tree, we can identify which features are most significant. For example, as seen below, points per game is the most important feature. Using this decision tree, we will identify the 3 - 5 most important features in building a playoff caliber team. For each of these features, we will then take one playoff team that

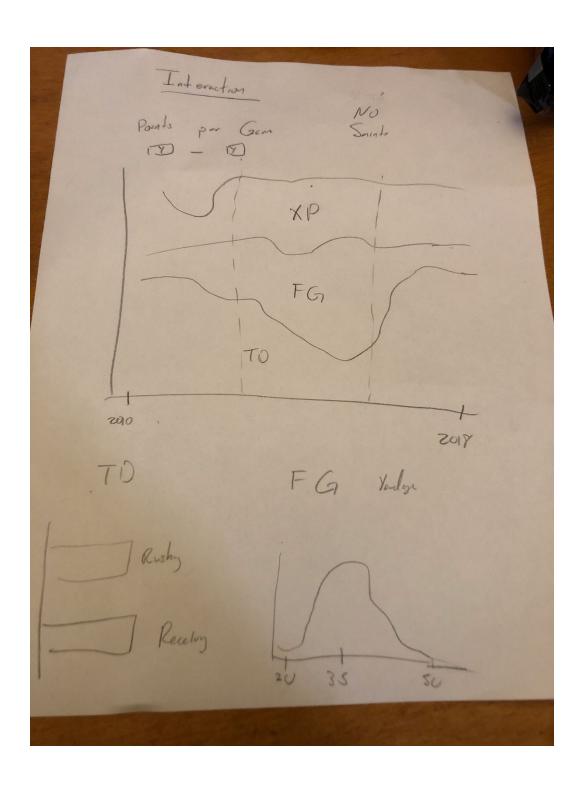
#### First design of an "innovative view" (decision tree)



This is a static picture representation of the decision tree that was built to understand what the hierarchy and effect of certain variables have upon the outcome of whether or not a team makes the playoffs. The final design will likely be more carefully pruned and involve a form of interaction, such as being able to click to show more nodes.

#### <u>Interactions (in the form of a storyboard or textual descriptions and some sketches)</u>

A sketch of our interaction is below. The important predictor of playoff success is number of points scored per page. To capture this, we use a stacked area chart, where each of the components of the stacked area chart is the number of points scored through the various means of scoring points. In this instance, the methods are from touchdowns, field goals, and extra points. We will chart this over time. The interaction functionality here is a brush that enables a user to select a specific time period. The graphs below the area chart will give more insight into the selected time frame, such as the proportion of passing / rushing touchdowns, or the distribution of field goal yardage.



## Prototype V2

#### **Decision Tree**

- Built a decision tree based on chosen predictors that included stats such as red zone performance, turnovers, offensive rushing, offensive passing, and defense
- <a href="https://www.kaggle.com/bhavesh09/titanic-decision-tree-visual-with-d3-js/notebook">https://www.kaggle.com/bhavesh09/titanic-decision-tree-visual-with-d3-js/notebook</a>
- https://bl.ocks.org/ajschumacher/65eda1df2b0dd2cf616f
- After building this decision tree, we identified four main categories that were nodes at the top of our decision tree
  - Each of these categories is identified below, with the accompanying highlighted team below

#### Points per Game

- Patriots 2010 - 2017

#### Opponent Offensive Yards

- 2015 Broncos
  - Scatter plot
    - Y axis total wins from the season
    - X axis offensive yards allowed

#### Turnover Differential

- 2015 Panthers
- TODO:
  - Turnover differential: defined as the number of interceptions and fumbles a team's defense has subtracted by the number of interceptions and fumbles its offense gets.
    - (Interceptions + Fumbles)\_{defense} (Interceptions + Fumbles)\_\_{offense}
  - Create a horizontal bar chart of 3 strong performing teams and 3 poorly performing teams in 2015
    - Vertical axis with horizontal bars

-

#### Rushing Yards

- 2014 Seahawks
- TODO:
  - Create 2 visualizations, 1 bar chart and 1 set of pie charts

We created visualizations for each of the categories chosen above.

#### **Narratives**

#### **Decision Tree**

- While various NFL teams and coaches hold different football philosophies on how to play the game, each organization ultimately has the same goal – to win games and make the playoffs. However, certain dynasties such as the Patriots have consistently made the playoffs in the last decade, teams like the Browns fail to make a single importance. To uncover the secrets of building a playoff caliber team, we collected data on all 32 teams from 2010 to 2017 and crafted a decision tree to see which features were critical to a winning team.
- "There are no shortcuts to building a team each season. You build the foundation brick by brick" William B Bellichick

#### Points per Game

- From 2010 2017, the New England Patriots dominated the league. They won the AFC East all 8 years, won the AFC championship 4 times, and won the Super Bowl 2 times. Unsurprisingly, they were a top 3 point scorer for 7 out of the 8 years.
- Compared with the entire NFL, the Patriots scored almost 50% more points every season with the same consistency as the average of 32 teams. The bulk of the additional points comes in the form of touchdowns, which stresses the advantage of touchdowns over field goals.

Quote: "the Patriots have been winning the same way every year: lots and lots of offense — with Tom Brady at the vanguard " - The Ringer (https://www.theringer.com/2016/7/28/16044166/nfl-most-consistent-teams-new-england-patriots-denver-broncos-seattle-seahawks-8a899d853621)

#### Defense

- 2016 Denver Broncos:

- In 2016, the Denver Broncos allowed significantly fewer yards than any other NFL team. Despite analysts clamoring over high-powered offenses such as Carolina and New England, the Denver Defense led the organization to a success 12 - 4 record. On their path through the playoffs, the defense held steadfast against both the Patriots and Panthers to secure the Lombardi trophy, with star linebacker Von Miller winning MVP honors.
- While points per game may be the strongest determinant of season success success,
- Bear Bryant is credited with the quote: "Offense sells tickets, but defense wins championships."
  - Fun fact this actually isn't true

#### Run game

- The NFL has emerged into a pass-heavy game, much more so than the history of the league has ever seen.

- "NFL records keep falling we're only 5 weeks into the 2018 season" said SBNation, as Tom Brady hit 500 career pass TD's and Drew Brees broke the all-time passing yards record.
- While passing records continue to fall with the incorporation of faster, spread offense, perhaps the run game is getting overlooked as the true key to success.
- The Seattle Seahawks are a prime example of this. The 2014 Seahawks burst into the playoff scene on the back of Pro Bowl running back Marshawn Lynch's 1306 rush yards and star guarterback Russell Wilson's ability to scramble for nearly 900 yards.
- Despite pundits and players' declarations that year that the NFL "is a passing league now", the Seahawks' rushing power led them all the way to a second consecutive Super Bowl appearance that year.

**Turnovers** 

- "Butler's interception comes off as a great play, which it really was, at the biggest time of the year"
  - Thomas Brady
- In Superbowl 49, Malcolm Butler iconically picked up Russell Wllson on the goal line to seal the Patriots' 24-21 victory over the Seahawks. Needless to say, big plays such as interceptions and fumbles can shift the momentum in a game and are huge determinants to the success of a team. In 2015, the Carolina Panthers led the league with a 15 1 record and won the NFC championships. Part of this success was driven by star all-pro cornerback Josh Norman and linebacker Luke Keuchly. The panthers were able to rack up a +26 turnover differential, with 24 picks verse 11 interceptions thrown and 28 fumble recoveries verse 15 lost. The Panthers dominated the NFL with their ability to take the ball away, and ultimately helped them land a spot in the Superbowl.

Tree:			
Stacked area chart:			

#### Scatter plot:

- Add the ability to scroll over years

#### Bar chart:

- When you scroll over the bar, highlight the row in the table

#### **Turnover bar chart:**

## Final Submission

This week, we met with Michael to discuss our final project from a main visualization standpoint, from a story telling standpoint, and from an additional visuals standpoint.

#### The comments from Michael a think-aloud exercise:

- 1. Strong main visualization two thumbs up
  - a. The great part of the visualization is the data mining behind the visual
  - b. Make sure that the text is readable
  - c. Implement some form of tool tip, thinks that it feels second nature to have
  - d. Overall fantastic, just make sure to clean up
- 2. Story telling
  - a. Story telling is nice, especially with the decision tree
  - b. Make sure it is clear from the title what the story we are telling
  - c. Flow of the visualizations make a lot of sense
  - d. Make sure to include an about us page at the end
- 3. Supplemental Visualizations
  - a. Make sure to include clear legends when needed and make sure color scheme is cohesive
  - b. Make sure it is obvious when to click on things
  - c. Just clean up things such as font so everything is cohesive
  - d. Animations visual is something that is really cool
  - e. Overall, supplemental visuals are also great and do a good job supporting the main story / main visualization

#### Addressing the comments:

- 1. Main visualization
  - a. Per Michael's comments, we abbreviate all the features to ensure that fit correctly
  - b. Included a tool tip that has information such as gini impurity, samples in node, and distribution
  - c. Made sure the arrows moved with the visuals
  - d. Made sure that the legend was clearly separated from the rest of the visual
- Story telling
  - a. Changed the title to underscore that the purpose of the project is "How to build a playoff caliber team"
  - b. Included additional supportive text to highlight the story
  - c. Included an about us page at the end
- 3. Supplemental Visualizations
  - a. Included the animations page

- b. Standardized all fonts
- c. Included legends to all visualizations
- d. Included labels to all axes for the visualizations
- e. Included clear instructions for visualizations (for example, click on bar to get data for the turnover differential visualization)