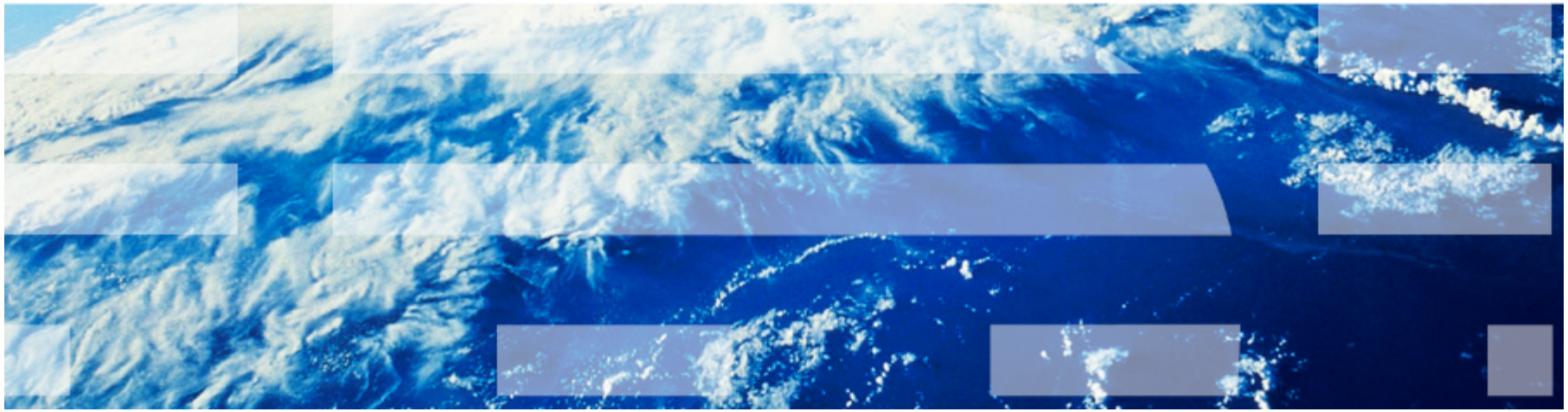
E6893 Big Data Analytics:

Final Project Presentation



Objective

- Classify Fantasy Football players as Boom or Bust or consistent scoring players
  - Boom or Bust: Scores a lot of points or very little points each week
  - Consistent: always scores around a similar amount of points each week, low variance in week to week scores
- Determine what makeup of Boom or Bust and Consistent players to give the team a best chance at winning over the course of a fantasy season

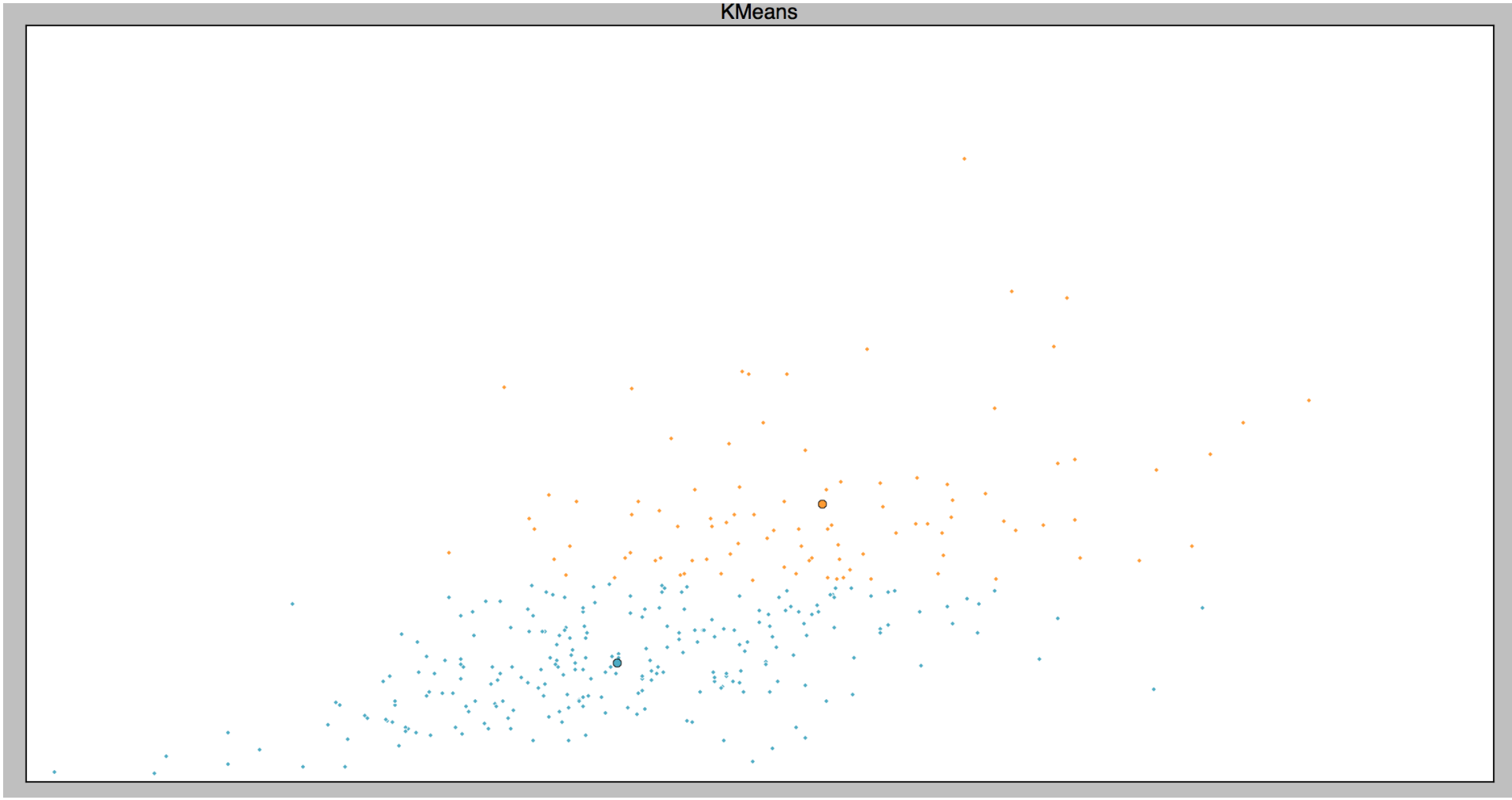
Setup, Data and Tools

- Python, scikit-learn, numpy, pyplot for graph data viewing
- Nfldb python package:
  - API and dataset for manipulating and analyzing NFL statistics
  - Stats from 2009-present day, trivial to update with recent data
  - Pulled by crawling the data on the NFL's official website, cached in postgresql DB
  - Very fine granularity; stats on specific players on specific plays
    - Which can actually make it a little complicated to get the aggregated statistics we want

Step One: Create tool for determining Boom/Bust vs Consistent Players

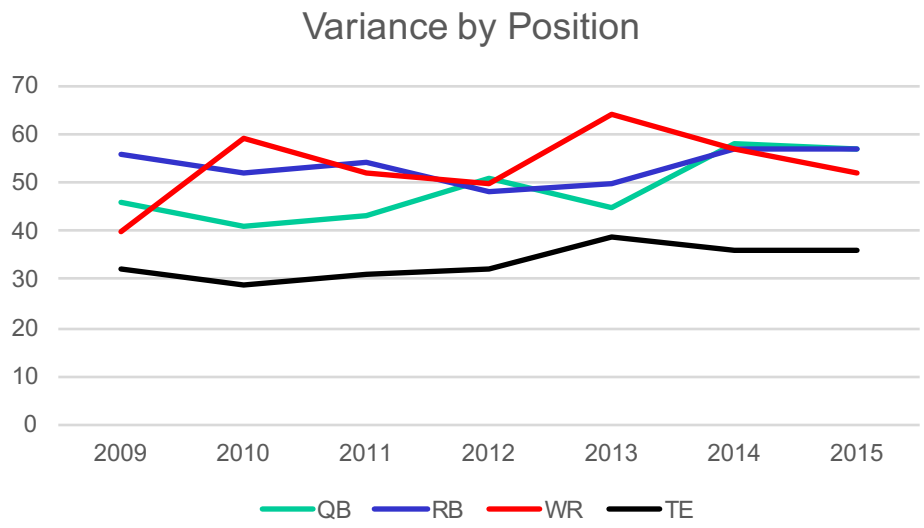
- Created a script for finding the top scoring 'fantasy relevant' players for each year for fantasy relevant positions (QB, RB, WR, TE)
  - This is usually the top 12 QBs (the average fantasy league has 12 or less teams) , top 24 RBs (2RBs/team), 36 WRs (3/team) and 12 TEs (1/team) each year
- Calculate their fantasy score for each game they played in the target year based on fantasy football 'Standard Scoring' rules
- Calculate various statistics on the list of scores for the player, such as variance, standard deviation, mean score, and quartile ranges
- Use the variance as a feature for the player and pass in the list of player/year combos to a KMeans clustering algorithm with k=2, the boom or bust player cluster and the consistent player cluster.
- We can now pass new players into the model to predict/classify them as a consistent or boom/bust type player

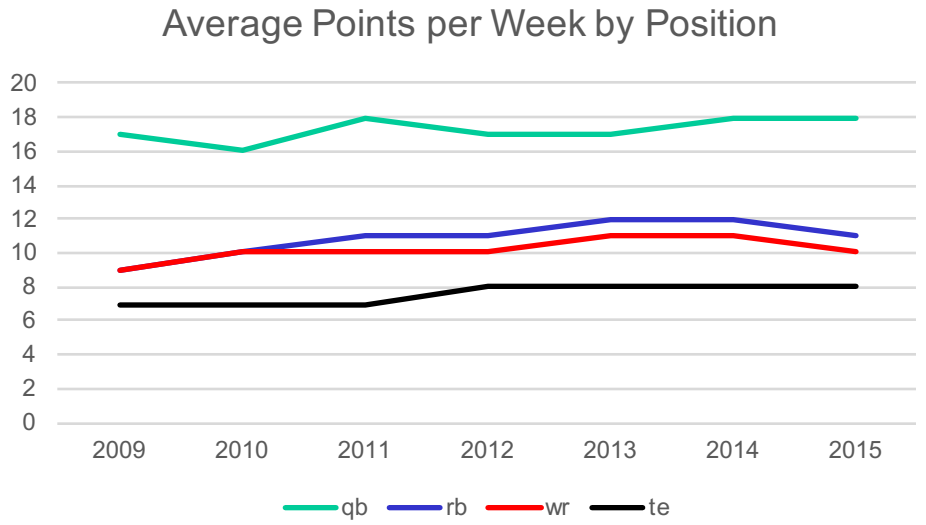
Graph of Clusters in KMeans Analysis

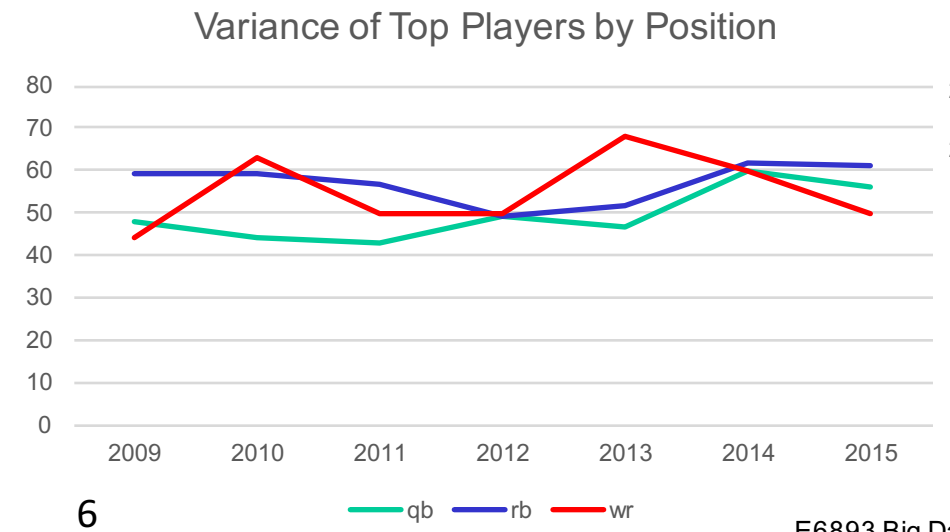


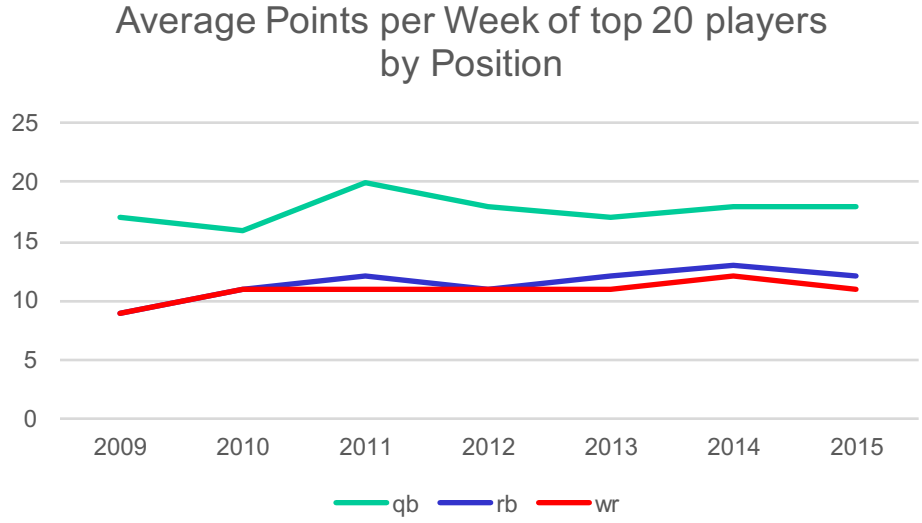
Orange is Boom/Bust and Blue is Consistent players

Positional Stats over Time





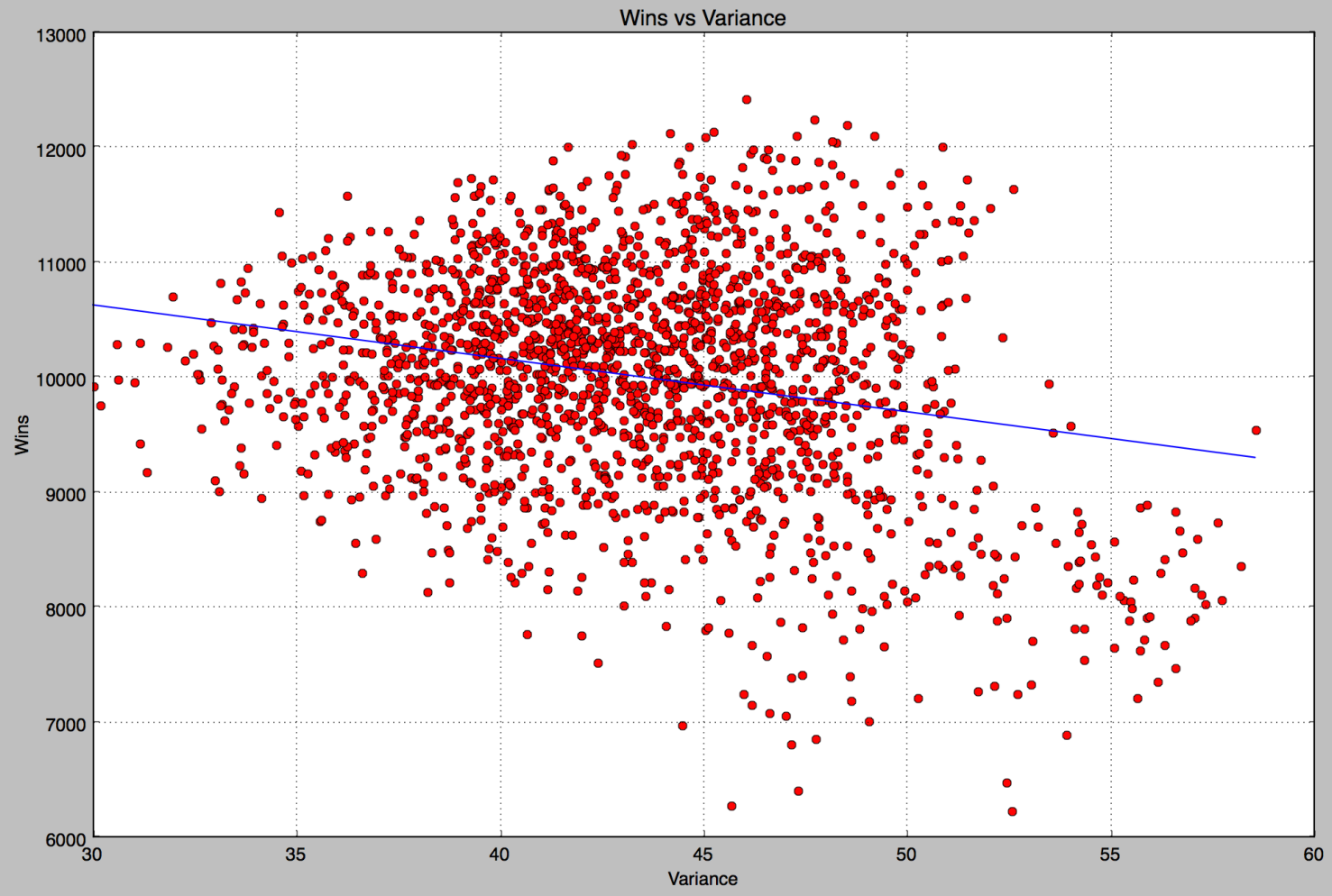




Analyzing Team Trends with Boom/Bust vs. Consistency

- From the previous analysis results an algorithm was created for finding 'average' teams (The mean score of the entire team was equivalent to the sum of the average score for each position over all positions)
  - Basically a modified version of the Knapsack Problem
- Next, we can simulate each team 'playing' each other team every week in the 'season' and record when one team outscores the other as wins and losses
- Then we can compare the average variance for the team to the wins for the team for each team and understand how variance affects the chance of a win

Team score trends and graph for it



Results

- Generally less variance leads to more wins
  - But a certain amount of variance is needed, as the highest wins were scored with a variance between 40-50
- Most teams clustered around a variance of 42
- Outliers: most of the worst and best teams had a variance between 45-50

Questions?