NHL Salary Analysis

Project Overview

Data: NHL data on players and teams.

Questions:

Can we predict player salaries based on stats?

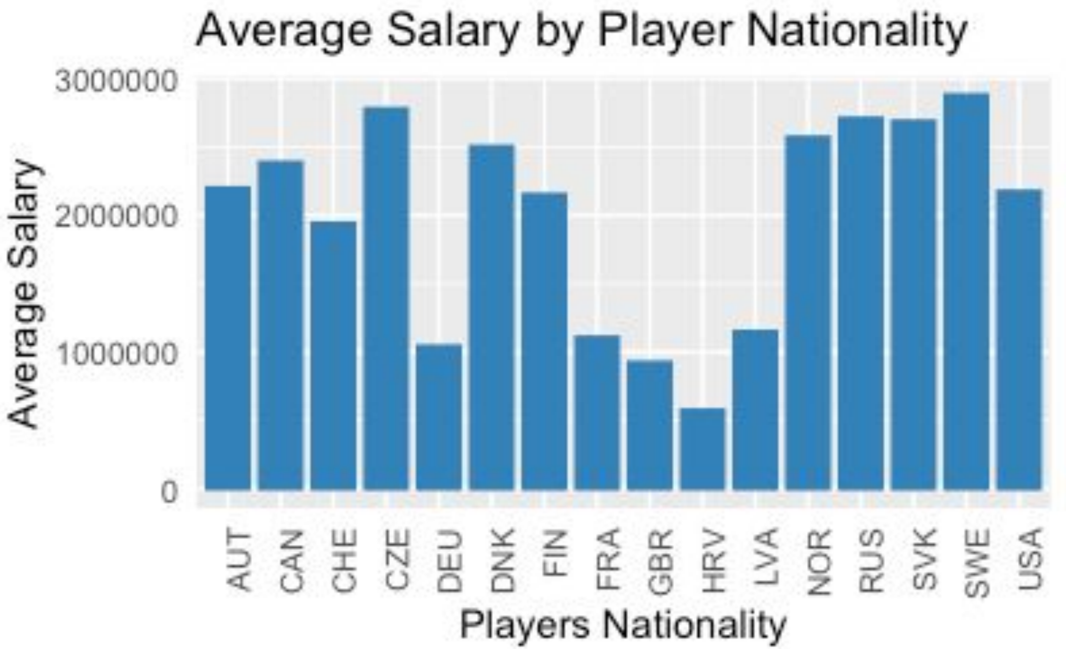
Should teams hire superstars?

Should the Vancouver Canucks hire Quinn Hughes?

Data Cleaning

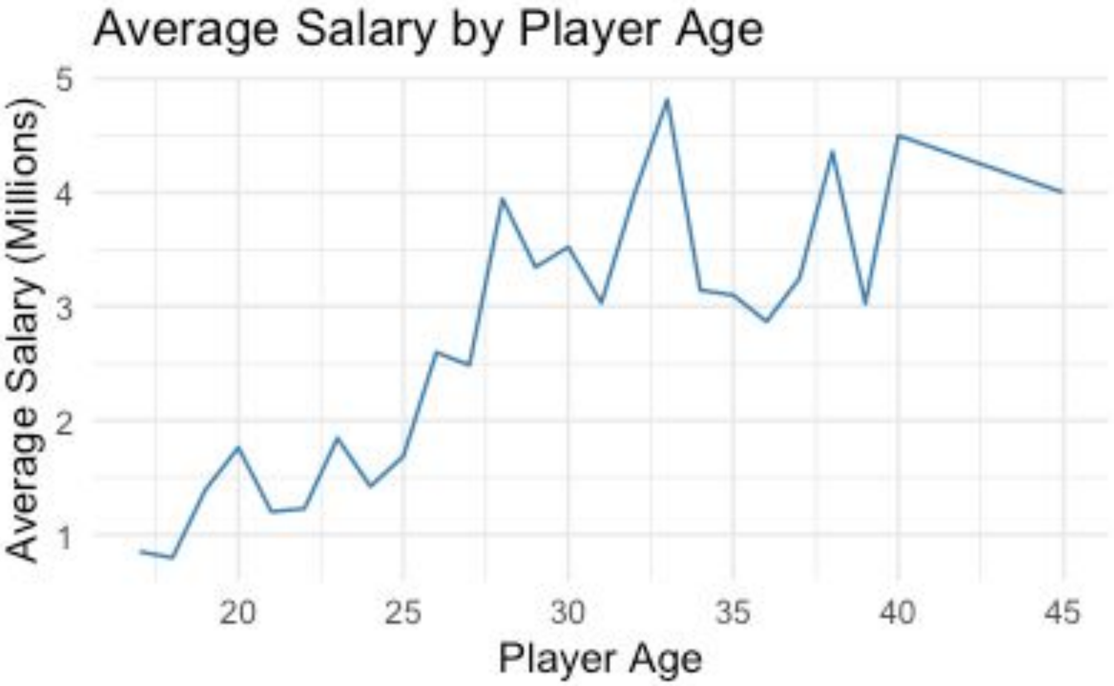
- Handling NA values
  - Providence/State
  - Draft Year, Round, and Overall
  - Average Shot Distance
  - Removal of players with high NA count
  - Replaced with mean values
- Remove Rookies
  - Matthews: 40 goals, 69 points
- Remove long-term injured players
  - Stamkos: 17 GP, $9.5 Million

Exploratory- Nationality



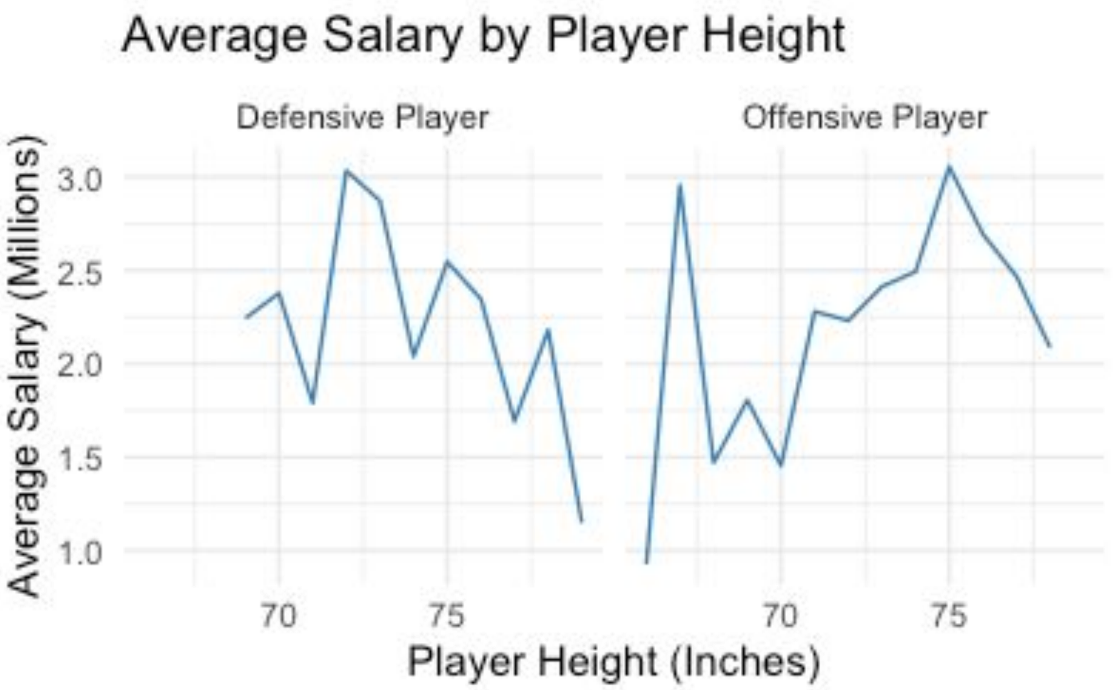
- The United State does not produce the highest paid Hockey players
- Highest paid players come from Europe
- USA/Canada are the only North American countries

Exploratory- Age



- A player's salary is likely to increase until their early 30's
- No dramatic dip in salary as age increases
- Peaks at age 33

Exploratory - Height



- Any two-way player is classified as an offensive player
- Offensive players can get away with being shorter
- Defensive players average salary peaks at a shorter height

Building the Model

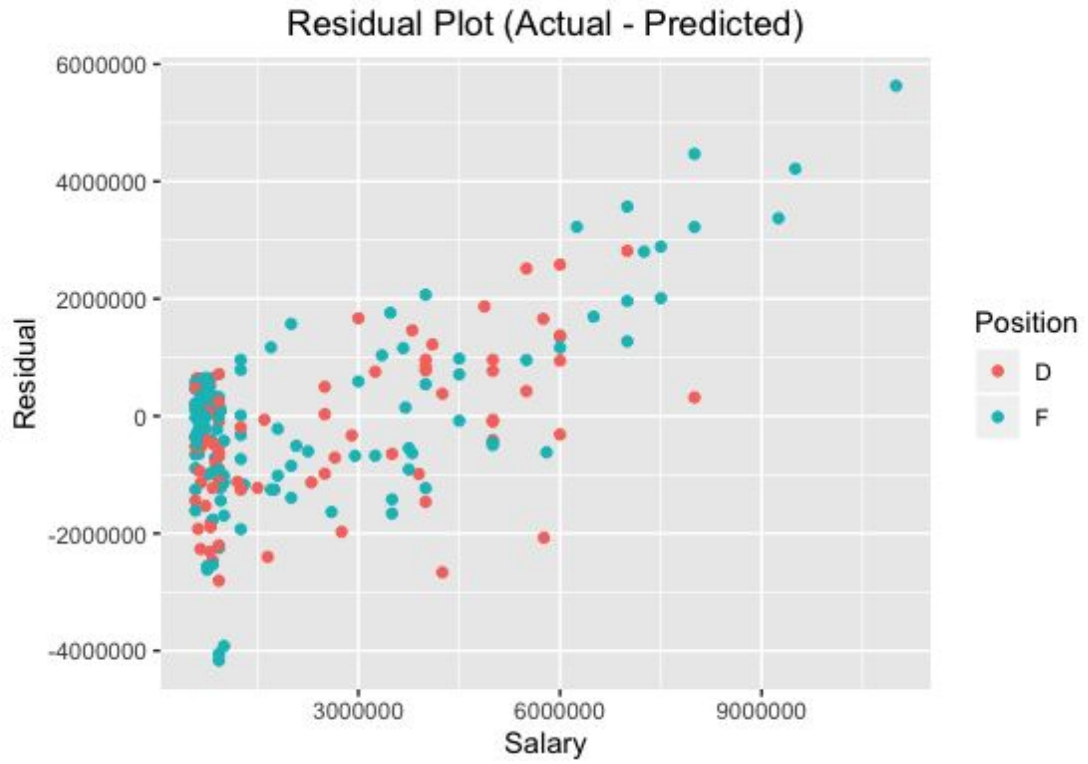
```r
set.seed(45)
cv_5 = trainControl(method = "cv", 5)

best_elastic_regression = train(
    form = Salary ~ .,
    data = stat_trn,
    method = "glmnet",
    trControl = cv_5,
    tuneLength = 10
  )
```

- Additional Data Cleaning
  - Categorical Data: Last Name, Country, City, etc
  - Face-Off Statistics
  - Double-counted statistics
  - Position → D and F only
- Penalized Regression
- Transformations attempted
  - Log, cube root, etc
- Lowest RMSE achieved:

  $1,436,393

Using the Model

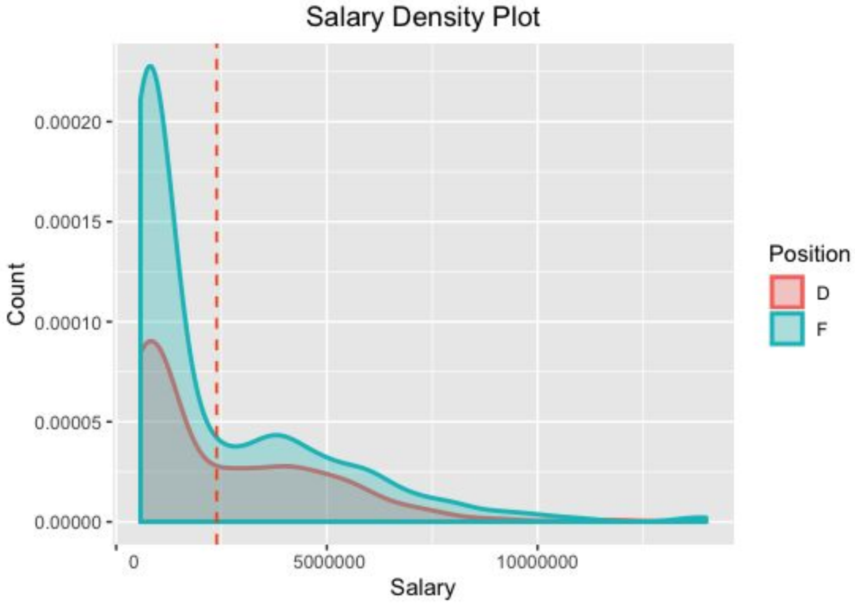


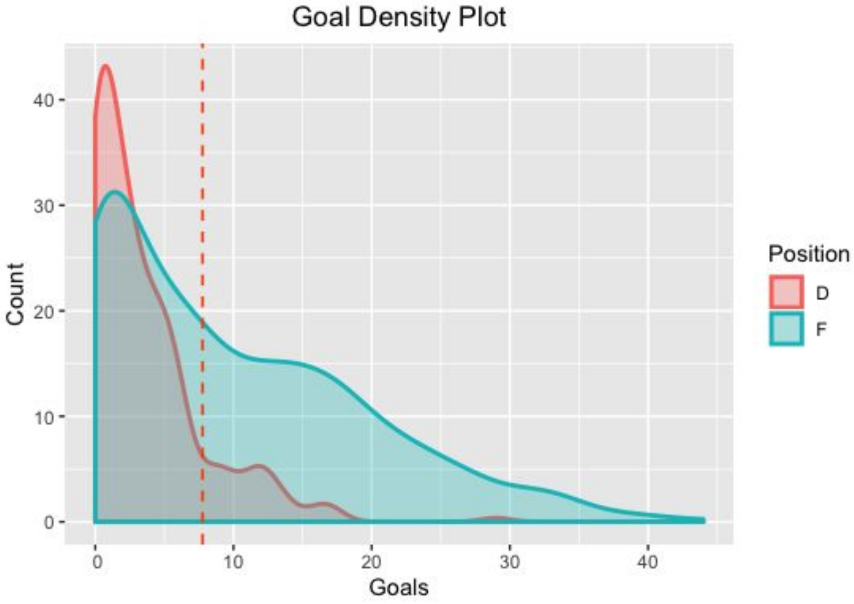
Quinn Hughes Predicted Salary:

$2,967,417

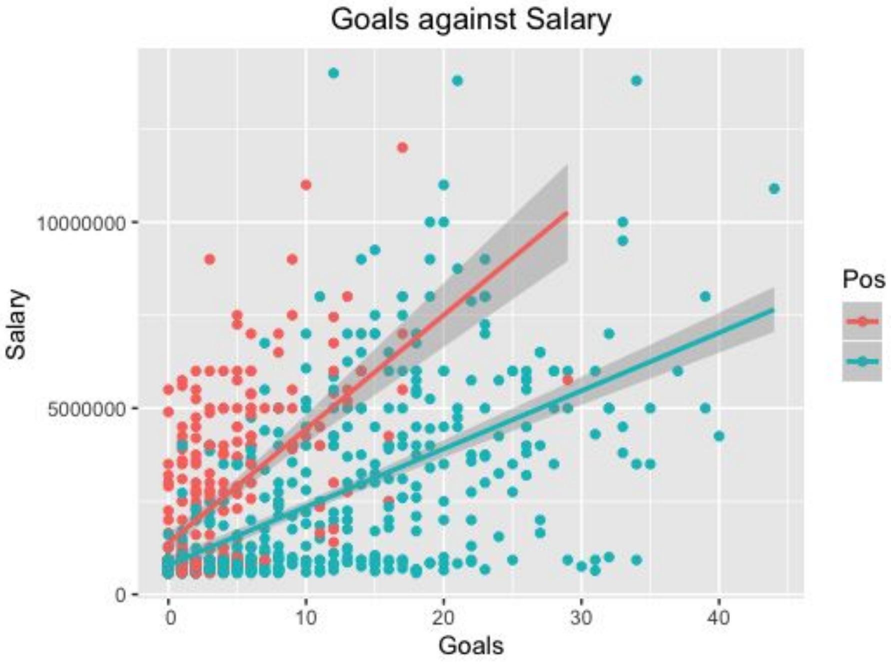
(We believe this is a huge underestimate)

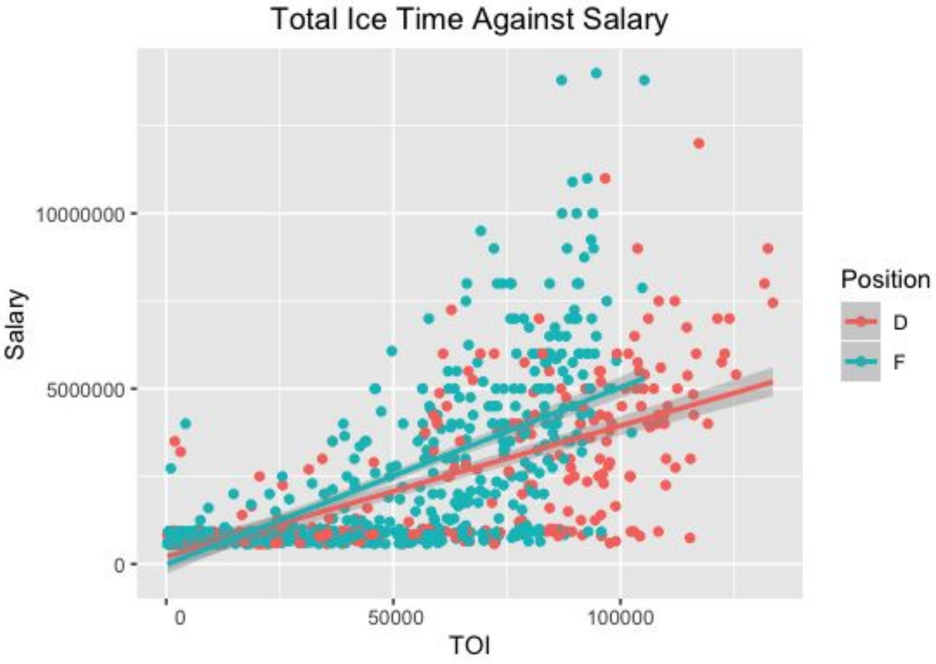
Reasons for Shortfall - Skewed Data





Reasons for Shortfall - Noisy Data

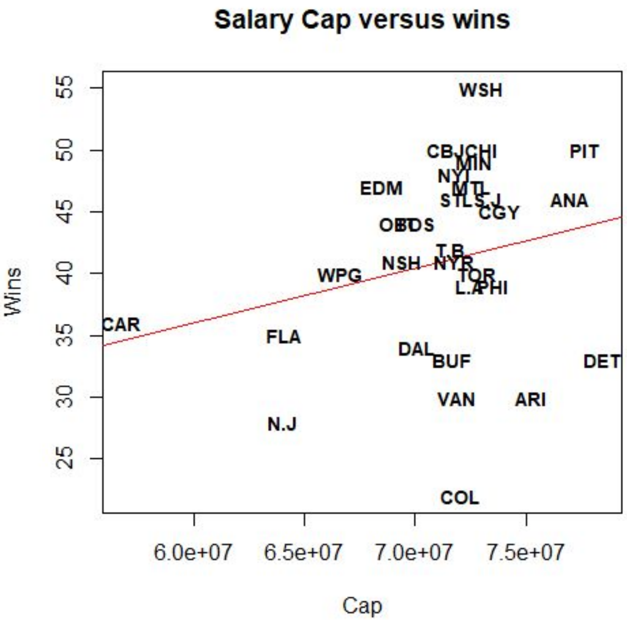


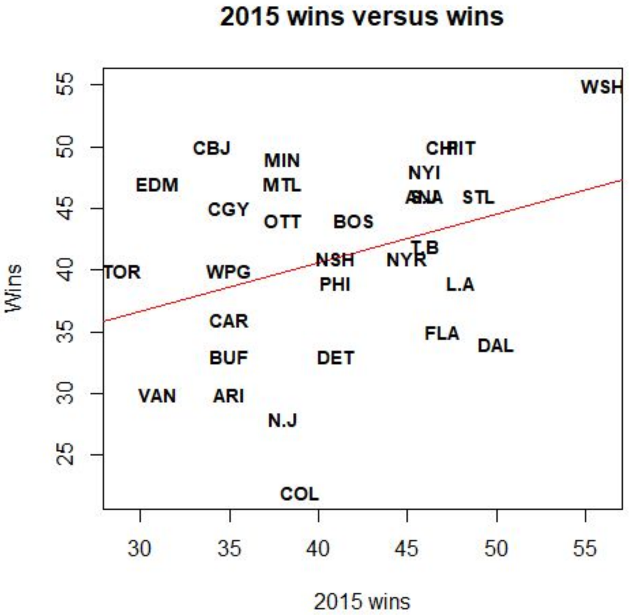


Directions for Improvement

- Increase Sample Size to multiple seasons
- Split data into baskets of similar playing styles
- Explore non-linear relationships further

Team Analysis

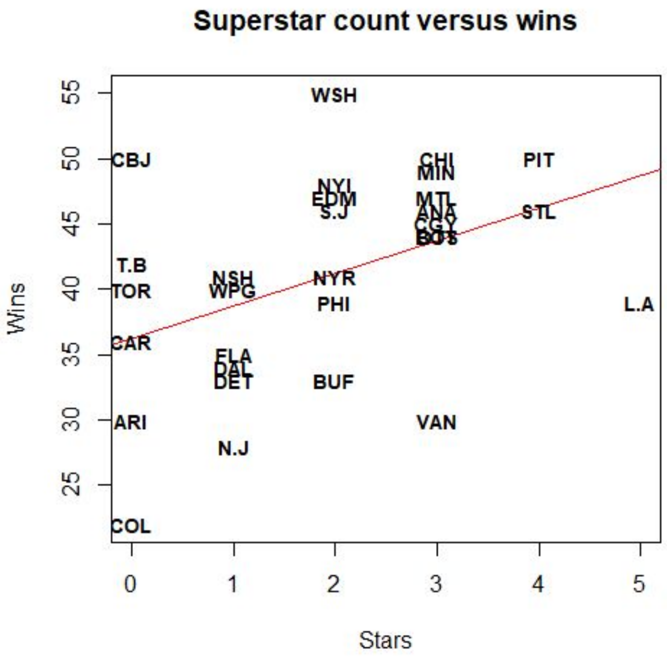


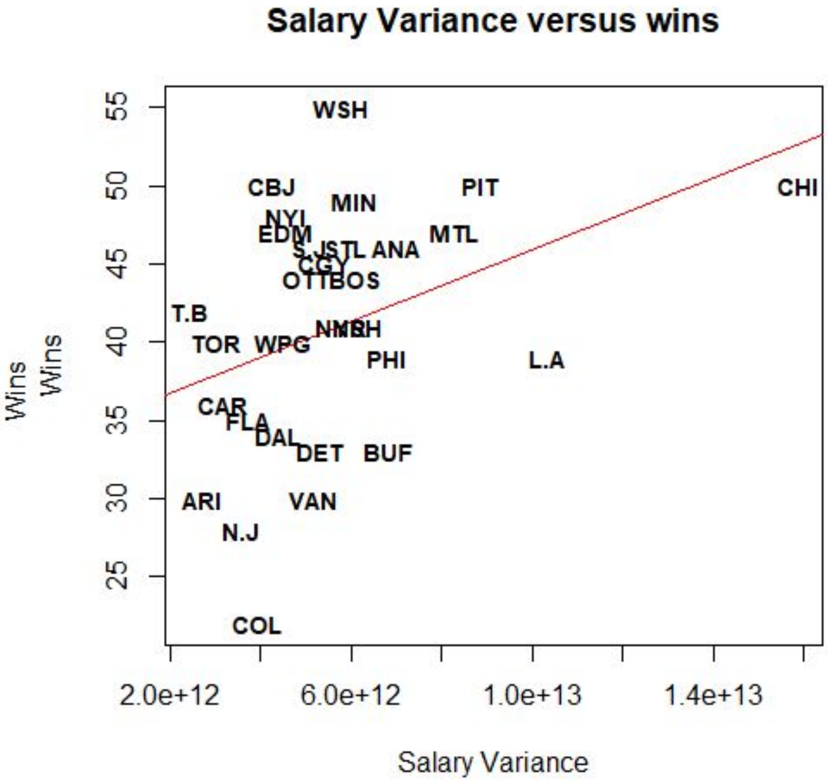


Regression

These two variables explain about 10-15 percent of the variance between teams.

Team Analysis





Regression

Bringing in these additional variables we are able to explain around 15-25 percent of the variance.

Each additional superstar is associated with 1.5 more wins.

Quinn Hughes

Difficulties and improvements

- Data Wrangling
- Time series
- Weaknesses in data

Questions?