Sports Analytics - The Story of Moneyball

Karishma Yadav

6/21/2020

The goal of this project is to study the story of a Baseball team - Oakland A's (San Francisco), which was one of the poorest teams in baseball when the owners introduced budget cuts in the year 1995. Despite this they were improving over the years 1997-2001 (refer to fig. 1 in Appendix), the percentage of wins kept on increasing over this time period. It was seen that they were winning as much as the rich teams that could afford all the star players (refer to fig. 2 in Appendix).

So how do the poor teams compete?

The Oakland A's followed a different approach. The traditional way of selecting the players was through scouting, the A's however selected the players based on statistics and not their looks. Quantitative analysis suggested that some skills were undervalued and some skills were overvalued. The key premise of Oakland A's was that if they could detect the undervalued skills, they could find players at bargain.

The goal of a baseball team is to make it to the playoffs, the A's approach was to use to analytics to derive a strategy to make it to the playoffs.

They figured out that they needed to figure out the number of games they need to win to make it to the playoffs and the number of runs they need to score to win a game. It was claimed by the team that the number of games to make is to the playoffs is 95 and number of runs scored to win 95 games should be 135 more than the runs allowed. We will confirm this using the data from the Baseball Reference website.

We start by loading the data, as follows:

```r
baseball <- read.csv("baseball.csv")
str(baseball)
```

```
## 'data.frame':    1232 obs. of  15 variables:
##  $ Team        : chr  "ARI" "ATL" "BAL" "BOS" ...
##  $ League      : chr  "NL" "NL" "AL" "AL" ...
##  $ Year        : int  2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 ...
##  $ RS          : int  734 700 712 734 613 748 669 667 758 726 ...
##  $ RA          : int  688 600 705 806 759 676 588 845 890 670 ...
##  $ W           : int  81 94 93 69 61 85 97 68 64 88 ...
##  $ OBP         : num  0.328 0.32 0.311 0.315 0.302 0.318 0.315 0.324 0.33 0.335 ...
##  $ SLG         : num  0.418 0.389 0.417 0.415 0.378 0.422 0.411 0.381 0.436 0.422 ...
##  $ BA          : num  0.259 0.247 0.247 0.26 0.24 0.255 0.251 0.251 0.274 0.268 ...
##  $ Playoffs    : int  0 1 1 0 0 0 1 0 0 1 ...
##  $ RankSeason  : int  NA 4 5 NA NA NA 2 NA NA 6 ...
##  $ RankPlayoffs: int  NA 5 4 NA NA NA 4 NA NA 2 ...
##  $ G           : int  162 162 162 162 162 162 162 162 162 162 ...
##  $ OOBP        : num  0.317 0.306 0.315 0.331 0.335 0.319 0.305 0.336 0.357 0.314 ...
##  $ OSLG        : num  0.415 0.378 0.403 0.428 0.424 0.405 0.39 0.43 0.47 0.402 ...
```

This dataset contains observations for every team and year pair for the years 1962-2012 for all seasons for 162 games. Here,

RA:Runs Allowed RS: Runs Scored W: Wins

Since the team Oakland A's used data till the year 2002, to predict for 2002, we will take the subset of our dataset for the same.

```
moneyball <- subset(baseball, Year<2002)
str(moneyball)
```

```
## 'data.frame':    902 obs. of  15 variables:
##  $ Team        : chr  "ANA" "ARI" "ATL" "BAL" ...
##  $ League      : chr  "AL" "NL" "NL" "AL" ...
##  $ Year        : int  2001 2001 2001 2001 2001 2001 2001 2001 2001 2001 ...
##  $ RS          : int  691 818 729 687 772 777 798 735 897 923 ...
##  $ RA          : int  730 677 643 829 745 701 795 850 821 906 ...
##  $ W           : int  75 92 88 63 82 88 83 66 91 73 ...
##  $ OBP         : num  0.327 0.341 0.324 0.319 0.334 0.336 0.334 0.324 0.35 0.354 ...
##  $ SLG         : num  0.405 0.442 0.412 0.38 0.439 0.43 0.451 0.419 0.458 0.483 ...
##  $ BA          : num  0.261 0.267 0.26 0.248 0.266 0.261 0.268 0.262 0.278 0.292 ...
##  $ Playoffs    : int  0 1 1 0 0 0 0 0 1 0 ...
##  $ RankSeason  : int  NA 5 7 NA NA NA NA NA 6 NA ...
##  $ RankPlayoffs: int  NA 1 3 NA NA NA NA NA 4 NA ...
##  $ G           : int  162 162 162 162 161 162 162 162 162 162 ...
##  $ OOBP        : num  0.331 0.311 0.314 0.337 0.329 0.321 0.334 0.341 0.341 0.35 ...
##  $ OSLG        : num  0.412 0.404 0.384 0.439 0.393 0.398 0.427 0.455 0.417 0.48 ...
```

Now we have 902 observations for the same 15 variables. Now we will build a linear regression model to predict the wins using the difference between the runs scored and runs allowed.

```
moneyball$rundiff = moneyball$RS - moneyball$RA
str(moneyball)
```

```
## 'data.frame':    902 obs. of  16 variables:
##  $ Team        : chr  "ANA" "ARI" "ATL" "BAL" ...
##  $ League      : chr  "AL" "NL" "NL" "AL" ...
##  $ Year        : int  2001 2001 2001 2001 2001 2001 2001 2001 2001 2001 ...
##  $ RS          : int  691 818 729 687 772 777 798 735 897 923 ...
##  $ RA          : int  730 677 643 829 745 701 795 850 821 906 ...
##  $ W           : int  75 92 88 63 82 88 83 66 91 73 ...
##  $ OBP         : num  0.327 0.341 0.324 0.319 0.334 0.336 0.334 0.324 0.35 0.354 ...
##  $ SLG         : num  0.405 0.442 0.412 0.38 0.439 0.43 0.451 0.419 0.458 0.483 ...
##  $ BA          : num  0.261 0.267 0.26 0.248 0.266 0.261 0.268 0.262 0.278 0.292 ...
##  $ Playoffs    : int  0 1 1 0 0 0 0 0 1 0 ...
##  $ RankSeason  : int  NA 5 7 NA NA NA NA NA 6 NA ...
##  $ RankPlayoffs: int  NA 1 3 NA NA NA NA NA 4 NA ...
##  $ G           : int  162 162 162 162 161 162 162 162 162 162 ...
##  $ OOBP        : num  0.331 0.311 0.314 0.337 0.329 0.321 0.334 0.341 0.341 0.35 ...
##  $ OSLG        : num  0.412 0.404 0.384 0.439 0.393 0.398 0.427 0.455 0.417 0.48 ...
##  $ rundiff     : int  -39 141 86 -142 27 76 3 -115 76 17 ...
```

We have added a new column called rundiff to the dataset moneyball.

```
WinsRM <- lm(W~rundiff, data=moneyball)
summary(WinsRM)
```

```
##
## Call:
## lm(formula = W ~ rundiff, data = moneyball)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2662  -2.6509   0.1234   2.9364  11.6570
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 80.881375   0.131157  616.67   <2e-16 ***
## rundiff      0.105766   0.001297   81.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.939 on 900 degrees of freedom
## Multiple R-squared:  0.8808, Adjusted R-squared:  0.8807
## F-statistic:  6651 on 1 and 900 DF,  p-value: < 2.2e-16
```

We see that both the intercept and the independent variable rundiff are significant and the r-squared is 0.88. This suggests that our model is strong to predict the wins given the runs scored and runs allowed. Now we will try to confirm the claim made at moneyball.

Our Regression Equation is :

W = 80.8813 + (0.1058)rundiff

for W >= 95, we need the RHS of the equation to also be >= 95.

This implies that the rundiff should be greater than 133.40, which confirms the claim that the run difference between runs scored and runs allowed should be greater than 135 to win atleast 95 games. Now we need to predict the number of runs a team will score which can be done using batting statistics and the number of runs they will allow, which be can be predicted using pitching and fielding statistics.

The A's discovered that 2 baseball statistics were more significant than anything else:

1. On Base Percentage (OBP): Percentage of time a player gets on base(including walks)
2. Slugging Percentage (SLG): How far a player gets around the bases on his turn (measures power)

Most teams and players in baseball focus on Batting Average(BA): getting the base hit by a ball. A's claim was that OBP and SLG were undervalued and BA was overvalued. We can use linear regression to verify which baseball statistics are more important to predict runs.

```
RunsRM <- lm(RS ~ OBP + SLG + BA, data=moneyball)
summary(RunsRM)
```

```
##
## Call:
## lm(formula = RS ~ OBP + SLG + BA, data = moneyball)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -70.941 -17.247  -0.621  16.754  90.998
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -788.46      19.70 -40.029  < 2e-16 ***
## OBP          2917.42     110.47  26.410  < 2e-16 ***
## SLG          1637.93      45.99  35.612  < 2e-16 ***
## BA           -368.97     130.58  -2.826  0.00482 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.69 on 898 degrees of freedom
## Multiple R-squared:  0.9302, Adjusted R-squared:   0.93
## F-statistic:  3989 on 3 and 898 DF,  p-value: < 2.2e-16
```

We see that all of our independent variables are highly significant and our r-squared is 0.93, which suggests a strong model.

However, we see that the coefficient of BA is negative which would imply that RS vary inversely with BA which is counter-intuitive. This tells us that this is a case of multi-collinearity i.e., the statistical parameters are highly correlated. We can check this as follows and then build a model without taking BA into consideration and see if it affects the strength of our model.

```
cor1 <- cor(moneyball$OBP, moneyball$BA)
cor2 <- cor(moneyball$SLG, moneyball$BA)
cor1;cor2
```

```
## [1] 0.8540549
```

```
## [1] 0.8140681
```

```
RunsRM <- lm(RS ~ OBP + SLG, data=moneyball)
summary(RunsRM)
```

```
##
## Call:
## lm(formula = RS ~ OBP + SLG, data = moneyball)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -70.838 -17.174  -1.108  16.770  90.036
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -804.63      18.92  -42.53   <2e-16 ***
## OBP          2737.77      90.68   30.19   <2e-16 ***
## SLG          1584.91      42.16   37.60   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.79 on 899 degrees of freedom
## Multiple R-squared:  0.9296, Adjusted R-squared:  0.9294
## F-statistic:  5934 on 2 and 899 DF,  p-value: < 2.2e-16
```

We see that the coefficients for OBP and SLG are still highly significant and the r-squared is about the same as before. This confirms the claims made at moneyball. Our Regression equation for

RS = -804.63 + 2737.77(OBP) + 1584.91(SLG) r-squared = 0.93

Similarly, we can build a model for Runs allowed using the following 2 independent variables:

1. Opponents On Base Percentage (OOBP)
2. Opponents Slugging Percentage (OSLG)

```
RunsAllowedRM <- lm(RA ~ OOBP + OSLG, data = moneyball)
summary(RunsAllowedRM)
```

```
##
## Call:
## lm(formula = RA ~ OOBP + OSLG, data = moneyball)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -82.397 -15.178  -0.129  17.679  60.955
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -837.38      60.26 -13.897  < 2e-16 ***
## OOBP         2913.60     291.97   9.979 4.46e-16 ***
## OSLG         1514.29     175.43   8.632 2.55e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.67 on 87 degrees of freedom
##   (812 observations deleted due to missingness)
## Multiple R-squared:  0.9073, Adjusted R-squared:  0.9052
## F-statistic: 425.8 on 2 and 87 DF,  p-value: < 2.2e-16
```

We get the regression equation as:

RA = -837.38 + 2913.60(OOBP) + 1514.29(OSLG) r-squared = 0.91

Now we will use our models to predict the number of games the 2002 Oakland A's will win. Since, we are predicting the outcomes before the games are played, we need to estimate the 2002 team statistics using the 2001 team statistics.

At the beginning of 2002 season, the Oakland A's had 24 batters on their roaster. Using the 2001 regular season statistics for these players, we get:

Team OBP is 0.339 Team SLG is 0.430

Similarly, we get OOBP and OSLG for 17 pitchers as 0.307 and 0.373 respectively.

Using the regression equations, we get:

```
RS <- -804.63 + 2737.77*(0.339) + 1584.91*(0.430)
RS
```

```
## [1] 804.9853
```

```
RA <- -837.38 + 2913.60*(0.307) + 1514.29*(0.373)
RA
```

```
## [1] 621.9254
```

```
rundiff <- RS-RA
Wins <- 80.8814 + 0.1058*(rundiff)
Wins
```

```
## [1] 100.2491
```

We find the following, for the year 2002:

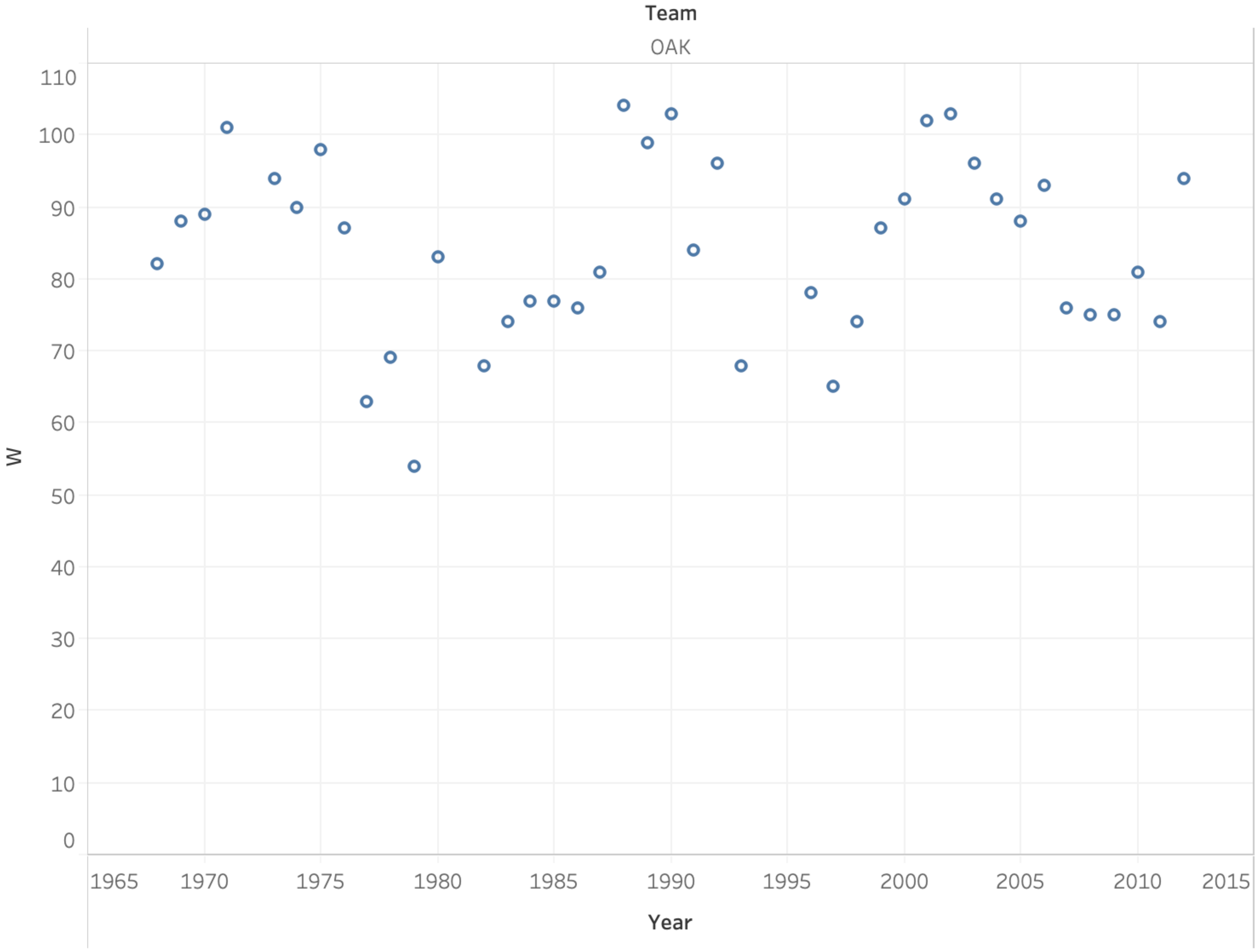
1. Oakland A predicted RS = 800-820, we predicted RS = 805, actual RS = 800
2. Oakland A predicted RA = 650-670, we predicted RA = 622, actual RA = 653
3. Oakland A predicted Wins = 93-97, we predicted Wins = 100, actual = 103

Therefore, our predictions closely match the actual performance. Hence, by using publically available data and analytics we can predict the outcome of a game before the game has even started. The correlation between the world series win and the playoffs is 0.3, i.e., winning regular season games gets one to the playoffs. but in the playoffs there are too few games foe luck to even out.

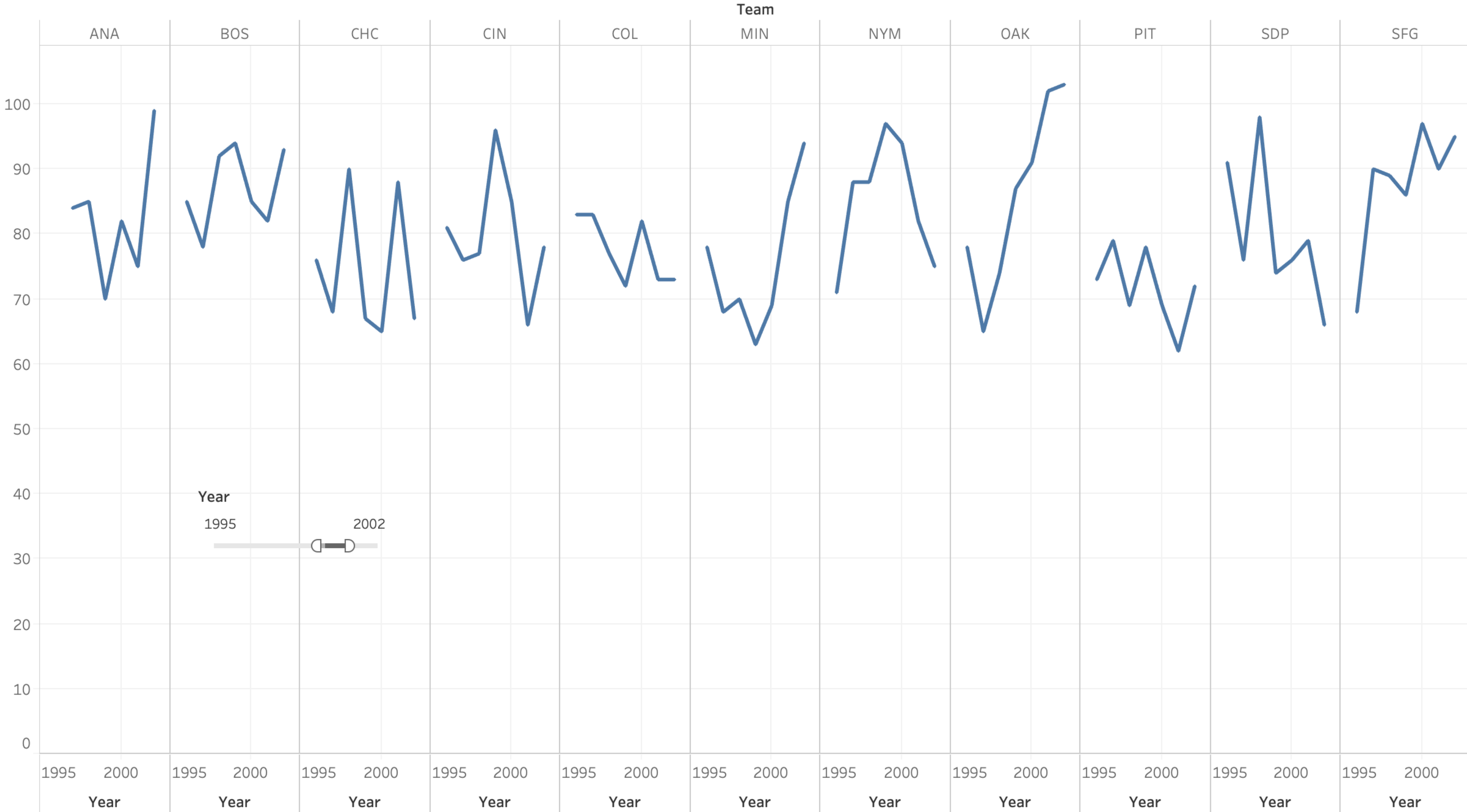
APPENDIX

The following figures have been created using Tableau.

Variation in number of games won by Oakland over the Years



Number of games won by Teams in Moneyball over the Years 1995-2002


Team
ANA
BOS
CHC
CIN
COL
MIN
NYM
OAK
PIT
SDP
SFG
Year
Year