StatR 101: Fall 2012

Final Project

Rod Doe

Monday, December 17, 2012

Table of Contents

[Analysis of Home Field Advantage in Major League Soccer 2012 1](#_Toc343457896)

[Data sources for this project 2](#_Toc343457897)

[The Significance of Home Field Advantage 2](#_Toc343457898)

[Team Summary Results 4](#_Toc343457899)

[Exploration of Correlations 6](#_Toc343457900)

[Time Zone Differential 6](#_Toc343457901)

[Travel Distance 10](#_Toc343457902)

[Great Circle Distance 10](#_Toc343457903)

[Back to Travel Distance 10](#_Toc343457904)

[Average Home Attendance 12](#_Toc343457905)

[Home Stadium Size 13](#_Toc343457906)

[Home Stadium Utilization 15](#_Toc343457907)

[Home Team Salary 17](#_Toc343457908)

[Conclusion 19](#_Toc343457909)

[Personal Notes 19](#_Toc343457910)

[Appendix A: R Code Used to Perform this Analysis 20](#_Toc343457911)

[Appendix B: Screen Scraper Script for www.goal.com 25](#_Toc343457912)

# Analysis of Home Field Advantage in Major League Soccer 2012

This is the report of a project that studied the effect of home field advantage in the 2012 Major League Soccer (MLS) season.

Project data reveal that in 341 matches, the home side won 174 matches, drew 85, and lost 82. Given that a team receives three points for a win, one point for a draw, and no points for a loss, the home teams accumulated 607 points, while the visitors accumulated 331 points.

Data studied include results of all MLS matches during the regular season and playoffs. The data do not include results of matches between MLS and non-MLS sides, nor do they include results from MLS sides playing in tournaments such as the US Open Cup or the Canada Cup. Put another way, the data include matches that contribute to MLS final standings, and the resultant MLS championship tournament for teams that qualified.

## Data sources for this project

Data sources for this project include:

|  |  |
| --- | --- |
| **Data** | **Source** |
| Home team, visitor, win, loss, draw, goals scored, goals allowed | [www.goal.com](http://www.goal.com) |
| Attendance | [www.mlsattendance.com](http://www.mlsattendance.com) |
| Travel Distances | Great circle distance calculated between cities coordinates. |

## The Significance of Home Field Advantage

The first impression is that these numbers are compelling, but what is the statistical evidence of a home field effect?

My null hypothesis is that home field advantage is a myth, and that one should expect the results of the matches played to be equally split between home team wins, draws, and home team losses.

H0: There was no advantage to play on home field in the 2012 MLS season.

Count of home wins = count of draws = count of home losses.

H1: There was an advantage to play on home field in the 2012 MLS season.

Count of home wins > count of draws and count of home wins > count of home losses.

Pearson’s chi-squared test quantifies the difference between the observed results and expected results. A contingency table with probabilities that sum to 1 is used to feed data to R.

|  |  |  |  |
| --- | --- | --- | --- |
| Team | Win  Observed  (Expected) | Draw  Observed  (Expected) | Lost  Observed  (Expected) |
| Home | 174  (113.67) | 85  (113.67) | 82  (113.67) |

The chi-squared statistic is calculated as ∑ (Observed – Expected)2 / Expected. With three data points, there are two degrees of freedom. (Two numbers can be whatever they want to be, but the responsible third number has to get the other two home safely and then be what is necessary to make them sum to 341. Last numbers are the designated drivers of the statistical world.) So, in this case, the chi-squared statistic works out to (174-113.67)2 / 113.67 + (85-113.67)2 / 113.67 + (82-113.67)2 / 113.67 = 48.08.

R of course makes this easier.

wldCount <- matrix(c(sum(mls$HomeWin), sum(mls$Draw), sum(mls$HomeLoss)), nrow=1)

wldCount

[,1] [,2] [,3]

[1,] 174 82 85

chisq.test(wldCount)

Chi-squared test for given probabilities

data: wldCount

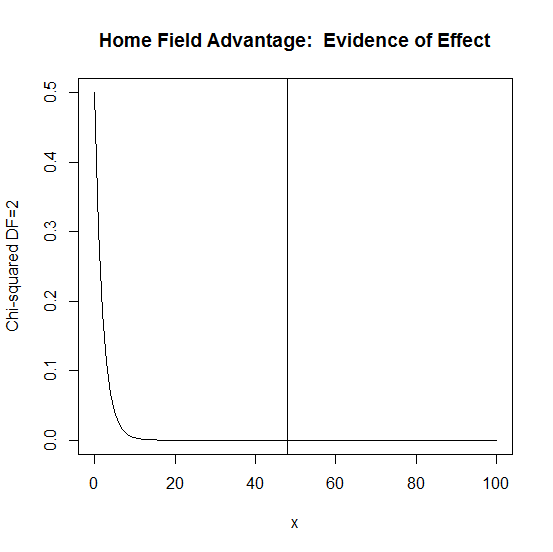
X-squared = 48.0762, df = 2, p-value = 3.634e-11

NOTE: Had the observed results equaled the expected results, the chi-squared statistic would have been 0, resulting in a p-value of 1, which is much larger than the α= 0.025 required for a 95% confidence level.

To plot this result, it is necessary to create a one-line function:

dwdl <- function(x) {dchisq(x, df=2)} #dwdl is chisq distribution of win-lose-draw

With this, the significance of this result can be appreciated graphically:



To drive this point home, let’s integrate this function to determine the area under the curve up to the chi-squared value of 48.08.

integrate(dwdl, 0, 48.08)

1 with absolute error < 4.8e-06

So there you have it – solid evidence of a home field advantage effect. These data allow rejection of the null hypothesis, and accept the alternate hypothesis that there was an advantage to play on home field in the 2012 MLS season.

## Team Summary Results

While most teams enjoy a home field advantage, not all do; some teams are clearly inferior. Chivas USA and Toronto FC were weak teams. Ironically, Toronto FC has the third highest total team salary in MLS.

NOTE: This report presents analysis of home field advantage as a function of salary. It’s an interesting result, so it appears near the end. Read on.

Group.1 HomeTeamScore VisitorScore

1 Chicago Fire 28 20

2 Chivas USA 9 30

3 Colorado Rapids 29 19

4 Columbus Crew 28 21

5 D.C. United 39 19

6 FC Dallas 21 16

7 Houston Dynamo 36 13

8 Los Angeles Galaxy 39 23

9 Montreal Impact 31 19

10 New England Revolution 23 15

11 New York Red Bulls 34 19

12 Philadelphia Union 22 23

13 Portland Timbers 24 21

14 Real Salt Lake 27 16

15 San Jose Earthquakes 44 25

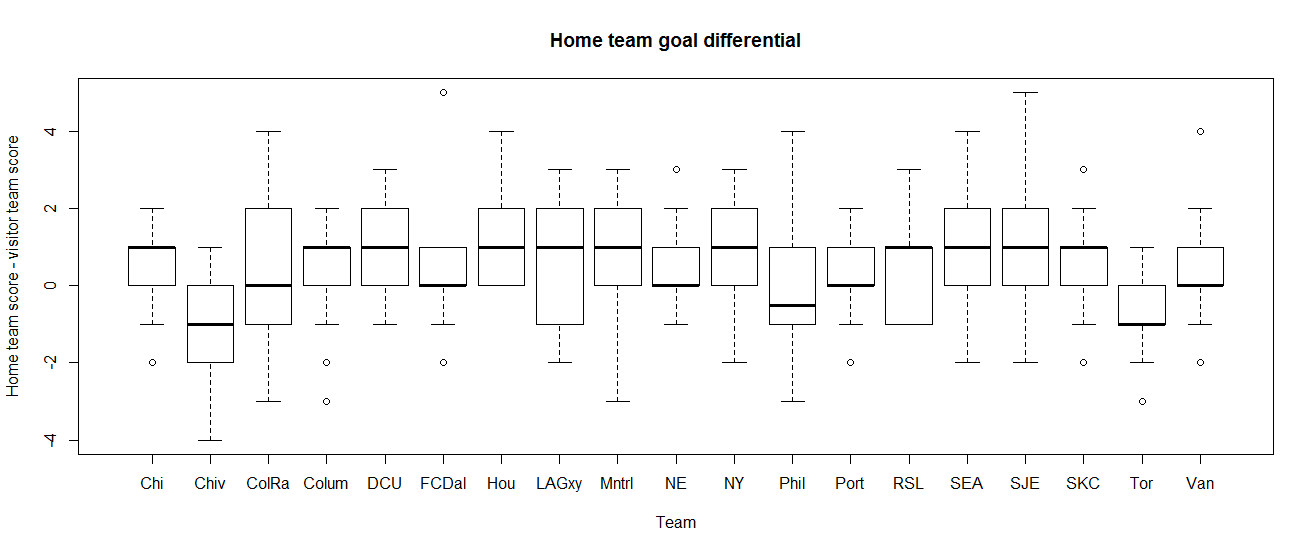
16 Seattle Sounders FC 29 12

17 Sporting Kansas City 24 13

18 Toronto FC 15 25

19 Vancouver Whitecaps 25 17

The home field advantage manifests itself as a goal differential between the home team and the visitor. Most teams enjoy a positive home team goal differential, but some do not. The home team goal differential shows teams’ home field advantages, or lack thereof.



# Exploration of Correlations

We will now explore attributes of the MLS season to find factors that correlate to home field advantage. These factors include:

* Visitor time zone differential
* Travel distance
* Home attendance
* Home stadium size
* Home stadium utilization
* Team salary

## Time Zone Differential

Given that MLS players range in age from the late teens to early thirties, and given the fact that these guys endure routine rigorous workouts, sleep should be something that affects performance. As such, travel across time zones could affect team performance. From a soccer perspective, this is a unique problem for MLS teams. The longest trip in MLS is from San Jose to New England (2328 nautical miles). This contrasts to the European experience in which the travel distances are much shorter.

Here are my hypotheses:

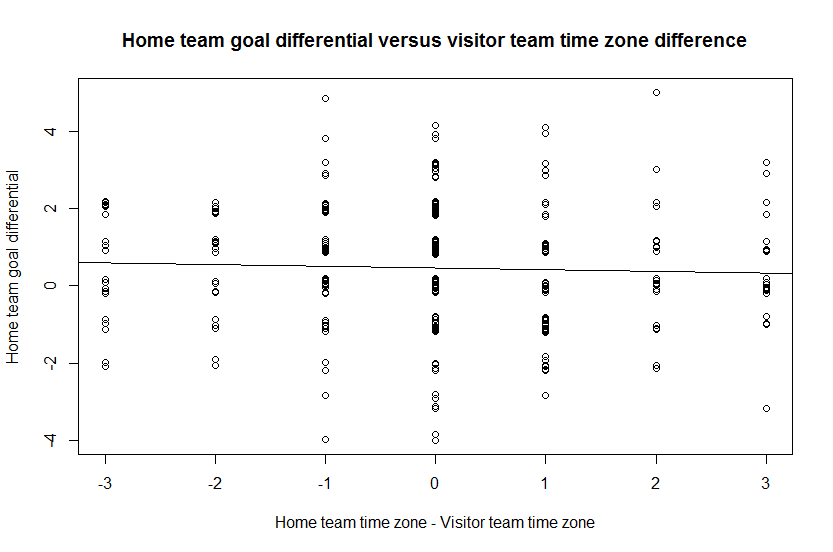
H0: Visitor team travel across time zones has no effect on home goal differential.

Mean home goal differential != f(visitor time zone differential)

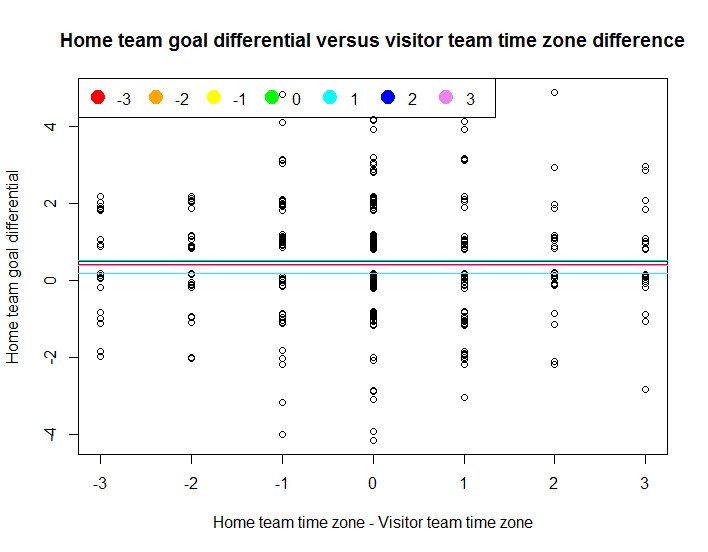
H1: There was an advantage to play on home against a team that traveled from a different time zone.

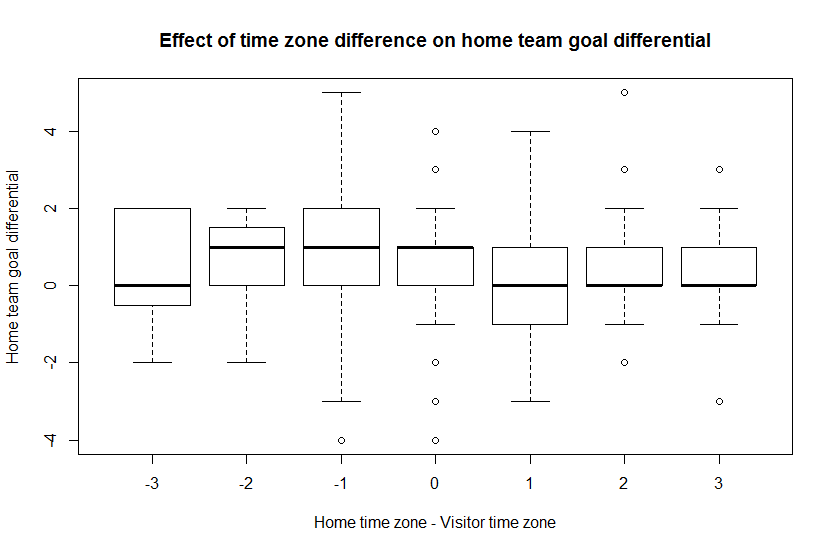
NOTE: In these data, west coast teams are in time zone 8 (offset from Greenwich), and east coast teams are in time zone 5. As such, when a west coast team plays on the east coast, the visitor time zone difference is 5 -8 = -3.

Here is a scatter plot of home goal differential versus visitor time zone difference:



The nearly flat abline through the cloud of points suggests that visitor time zone difference does not have much effect on home field advantage. This plot adds lines that plot the means for each time zone difference, to highlight the lack of effect.



The boxplot of these data is similarly unspectacular.

The R function summary.lm yields a numeric confirmation of what the plots suggest.

Call:

aov(formula = mls$HomeGoalDiff ~ mls$VisitorTimeZoneDiff)

Residuals:

Min 1Q Median 3Q Max

-4.5149 -0.6003 0.4424 0.6133 4.6133

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.47214 0.08276 5.705 2.54e-08 \*\*\*

mls$VisitorTimeZoneDiff -0.04274 0.05768 -0.741 0.459

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.528 on 339 degrees of freedom

Multiple R-squared: 0.001617, Adjusted R-squared: -0.001328

F-statistic: 0.549 on 1 and 339 DF, p-value: 0.4593

There is insufficient evidence to reject the null hypothesis. As such, visitor time zone traversal has no effect on home field goal differential. This surprised me.

## Travel Distance

We demonstrated that time zone traversal has no effect on home goal differential. Given that, we probably will not find a correlation between travel distance and home goal differential. However, let’s keep an open mind.

It should be noted that LA Galaxy and Chivas USA play in the same stadium. One would think that there is no home team advantage when those teams play, but as it turns out, the stadium is configured differently for the Galaxy than it is for the Goats. As such, I chose to not treat the situation in which Galaxy and Chivas play as a special case.

H0: Visitor team travel distance has no effect on home goal differential.

Mean home goal differential != f(visitor travel distance)

H1: There was an advantage to play at home against a team that traveled from a distance.

### Great Circle Distance

A great circle between two points is formed by the intersection of the surface of a sphere with a plane that passes through both points and the center of the sphere. In spite of its name, the great circle is the *shortest* distance between the two points, and for this reason, is used in navigation. In this report, the travel distance was obtained by calculation of the great circle distance between city coordinates. This equation to do this maybe not common knowledge, but it is quite handy. Here is the Excel formula used to calculate this value:

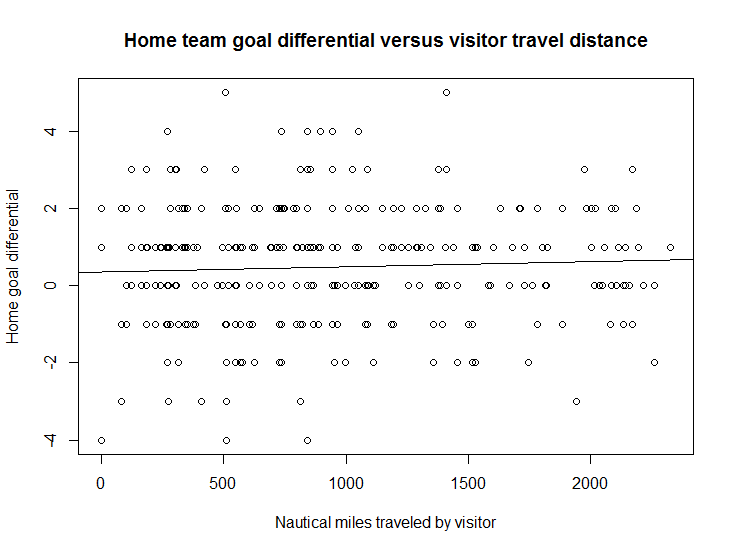
=60\*DEGREES(ACOS(SIN([@HomeLat])\*SIN([@VisitorLat]) +COS([@HomeLat])\*COS([@VisitorLat])\*COS(ABS([@HomeLong] -[@VisitorLong]))))

All of the coordinates must be expressed in radians.

NOTE: The author is a graduate of the U.S. Merchant Marine Academy, who obtained a Master’s license before coming ashore for a civilized life. He had no idea that he would ever use this formula again.

### Back to Travel Distance

The linear regression line through the cloud of points on the scatter plot of the home goal differential as a function of travel distance shows a positive correlation.



While the effect is apparent, it is insufficient to reject the null hypothesis.

Call:

aov(formula = mls$HomeGoalDiff ~ mls$NauticalMiles)

Residuals:

Min 1Q Median 3Q Max

-4.4646 -0.6665 0.3544 0.6236 4.5831

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.3447709 0.1510079 2.283 0.023 \*

mls$NauticalMiles 0.0001423 0.0001411 1.008 0.314

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.527 on 339 degrees of freedom

Multiple R-squared: 0.002989, Adjusted R-squared: 4.771e-05

F-statistic: 1.016 on 1 and 339 DF, p-value: 0.3141

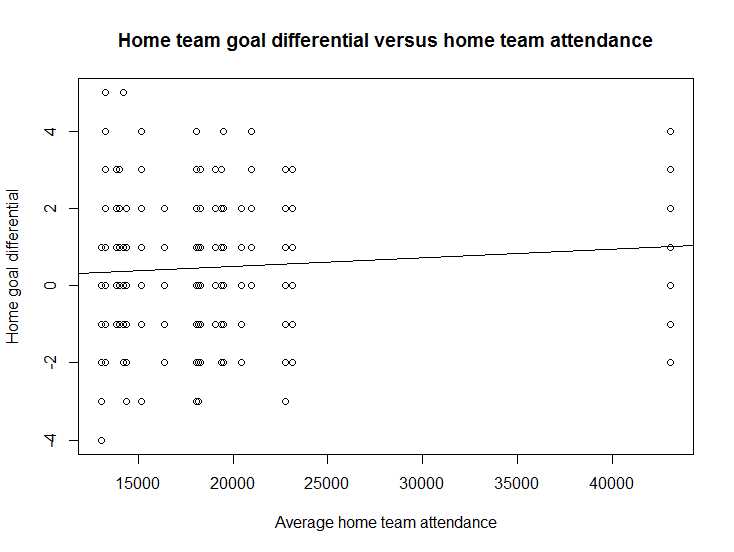
## Average Home Attendance

We fans like to think that we can affect the outcome of a match. We scream and yell and taunt, and in our beer-addled minds think that our taunts actually cause a player like Landon Donovan to whimper internally, “They don’t like me.” Let’s see what the numbers say about that.

H0: Home team crowd size has no effect on home goal differential.

Mean home goal differential != f(home team attendance)

H1: The home team realizes a statistically significant advantage when it plays before a large crowd.



The first thing to note is that the graph, while odd-looking, is correct. Moreover, we in Seattle are responsible for this odd appearance. The MLS average attendance is 18,800, but we in Seattle are over 43,000. No other MLS team even comes close.

The linear regression line indicates that there is a positive correlation between home goal differential and average attendance. But is it enough to reject the null hypothesis?

Call:

aov(formula = mls$HomeGoalDiff ~ mls$HomeTeamAttendance)

Residuals:

Min 1Q Median 3Q Max

-4.3380 -1.0200 -0.0200 0.6566 4.6566

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.164e-02 2.494e-01 0.167 0.8675

mls$HomeTeamAttendance 2.270e-05 1.241e-05 1.829 0.0683 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.522 on 339 degrees of freedom

Multiple R-squared: 0.009773, Adjusted R-squared: 0.006852

F-statistic: 3.346 on 1 and 339 DF, p-value: 0.06826

This p-value is so close to the alpha level for a one-sided 95% confidence interval. It is close to significant, but not quite there yet. Perhaps when San Jose builds their new stadium, they will push this number over the threshold. For now, there is insufficient evidence to reject the null hypothesis. Sorry, 12th Man. Have another beer.

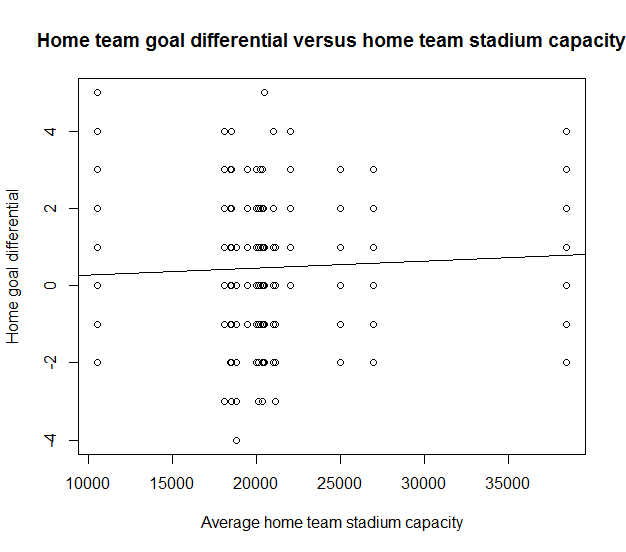
## Home Stadium Size

This one was a little flippant. Does stadium size matter?

H0: Stadium seating capacity has no effect on home goal differential.

Mean home goal differential != f(Stadium seating capacity)

H1: The home team realizes a statistically significant advantage when it plays in a large stadium.



Call:

aov(formula = mls$HomeGoalDiff ~ mls$HomeTeamCapacity)

Residuals:

Min 1Q Median 3Q Max

-4.4303 -0.7768 0.2232 0.5755 4.7152

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 9.972e-02 3.431e-01 0.291 0.771

mls$HomeTeamCapacity 1.759e-05 1.572e-05 1.119 0.264

Residual standard error: 1.527 on 339 degrees of freedom

Multiple R-squared: 0.003677, Adjusted R-squared: 0.0007379

F-statistic: 1.251 on 1 and 339 DF, p-value: 0.2641

Again, there is insufficient evidence to reject the null hypothesis.

## Home Stadium Utilization

This one caught my interest. Is it better to play in a large, empty stadium, or a small, packed house?

The utilization statistic is calculated by dividing the attendance by the stadium capacity. The results are quite surprising.

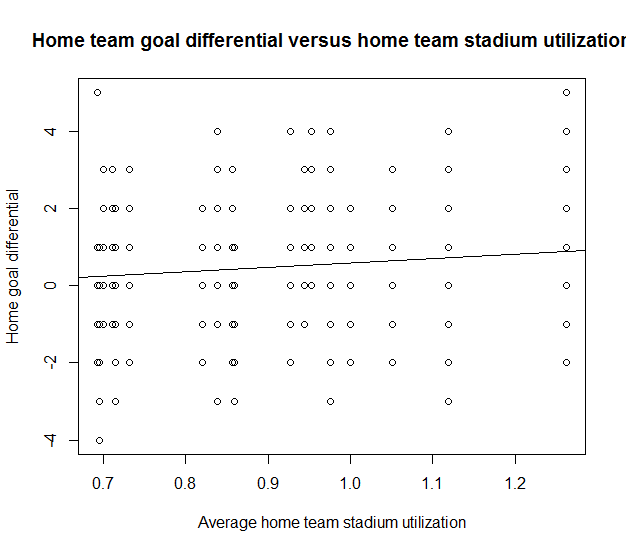
In MLS, there are some stadia that are routinely packed. Some of the stadia are overloaded:

|  |  |
| --- | --- |
| **Team** | **Utilization** |
| San Jose | 1.27 |
| Seattle | 1.12 |
| Montreal | 1.12 |
| Kansas City | 1.05 |
| Portland | 1.0 |

H0: Stadium seating utilization has no effect on home goal differential.

Mean home goal differential != f(Stadium seating utilization)

H1: The home team realizes a statistically significant advantage when it plays in a full stadium.



This effect is statistically significant, with a p-value sufficient to overturn the 95% confidence interval.

Call:

aov(formula = mls$HomeGoalDiff ~ mls$FractionFull)

Residuals:

Min 1Q Median 3Q Max

-4.2448 -0.8868 0.1132 0.7552 4.7573

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.5394 0.4582 -1.177 0.2400

mls$FractionFull 1.1292 0.5032 2.244 0.0255 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.518 on 339 degrees of freedom

Multiple R-squared: 0.01463, Adjusted R-squared: 0.01173

F-statistic: 5.035 on 1 and 339 DF, p-value: 0.02549

With 95% confidence, I reject that stadium utilization has no effect on home goal differential. We learn that a packed small stadium is better than a large, less densely packed house.

## Home Team Salary

Now we learn if money can buy success. It has been observed that market forces are not effective in sports. In the book Moneyball by Michael Lewis, the author notes that during one baseball season, the standings in the American League West were inversely proportional to team salary. It is interesting to note that it was the Seattle Mariners had the highest payroll and finished last.

In MLS, there is considerable salary disparity. The big spenders are New York Red Bulls ($12.96M), LA Galaxy ($12.63M), and Toronto FC ($8.25M). Here is what all that money bought them:

* New York was eliminated in the first round of the playoffs.
* LA won the MLS cup.
* Toronto had the second-worst record in MLS.

The bottom spenders are San Jose Earthquakes ($3.21M), Sporting Kansas City ($3.12M), Montreal Impact ($3.03M), and Houston Dynamo ($3.0M). Here is how they suffered:

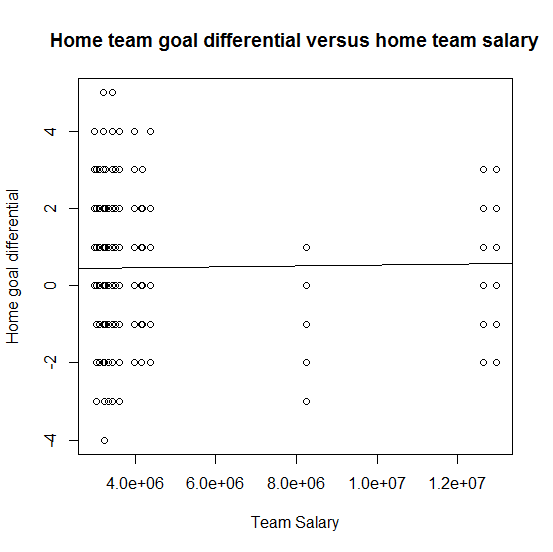
* San Jose finished on top of the MLS western conference, and was eliminated in the first round of the playoffs.
* Sporting Kansas City finished on top of the MLS east, and was eliminated in the first round of the playoffs.
* Montreal finished in the middle of the pack in their first season.
* Houston finished fifth in the MLS east, and advanced to the MLS championship match as a wildcard, losing to the LA Galaxy.
  + Houston had the highest home goal differential in MLS. When a match started in Houston, the score was essentially 1.2 – 0 in their favor.

Here are the hypotheses:

H0: Team salary has no effect on home goal differential.

Mean home goal differential != f(team salary)

H1: The home team realizes a statistically significant advantage when it spends more on team salary.



The linear regression line is almost flat.

Call:

aov(formula = mls$HomeGoalDiff ~ mls$HomeTeamSalary)

Residuals:

Min 1Q Median 3Q Max

-4.4554 -0.5588 0.4412 0.5467 4.5448

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.211e-01 1.537e-01 2.739 0.00648 \*\*

mls$HomeTeamSalary 1.062e-08 2.696e-08 0.394 0.69390

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.529 on 339 degrees of freedom

Multiple R-squared: 0.0004575, Adjusted R-squared: -0.002491

F-statistic: 0.1552 on 1 and 339 DF, p-value: 0.6939

There is insufficient evidence to reject the null hypothesis. Evidently, it is not possible to buy a home field goal differential.

# Conclusion

There is statistically significant evidence of a home field advantage effect in Major League Soccer. Additionally:

* Visitor time zone differential, to my surprise, had no significant effect on home field advantage.
* Travel distance had no significant effect on home field advantage.
* Home attendance had no significant effect on home field advantage, but may soon.
* Home stadium size had no significant effect on home field advantage.
* Home stadium utilization had a significant effect on home field advantage.
* Team salary had no effect on home field advantage.

After my assumptions failed, I did some cursory web research on factors attributed to home field advantage in numerous sports. The sources also identified travel as being insignificant. One source indicated that testosterone levels are higher at home (for whatever reason), and was thought to be a contributing factor. It would be useful to research if home field advantage has lower significance in women’s sports.

# Personal Notes

The time spent gathering and massaging data was about 5X as much as the time spent to research and present the data. If I had not written the PowerShell script (or the equivalent in Perl or Python or whatever) to fix the data culled from www.goal.com, I would still be untangling data.

This was a fun project. It challenged some core beliefs and taught me to dismiss personal biases and to just follow the data. And R just rocks! ☺

# Appendix A: R Code Used to Perform this Analysis

setwd("C:/Users/Rod/SkyDrive/R/101/Project")

mls<-read.csv(file="MLSWinDrawLose2012.csv")

# Q: Home field advantage - is it significant in MSL soccer?

# H0: There was no such thing as home field advantage in the 2012 MLS soccer season.

# Expressed mathematically:

# Count of home wins = count of draws = count of home losses.

# H1: It was advantageous to play on the home field in the 2012 MLS soccer season.

# Expressed mathematically:

# Count of home wins > count of draws and count of home wins > count of home losses.

# Chi-squared test of counts of home wins, home draws, and home losses.

# Nuances: Exclude results from LA Galaxy versus Chivas USA, since they both play in the same stadium.

# Summarize home team advantage in wins.

data.frame(sum(mls$HomeWin), sum(mls$Draw), sum(mls$HomeLoss))

# But wait, it's even worse. Summarize home points.

data.frame(sum(mls$HomePoints), sum(mls$VisitorPoints))

# Get win-draw-loss sums into a matrix, then perform chisq-test.

wldCount <- matrix(c(sum(mls$HomeWin), sum(mls$Draw), sum(mls$HomeLoss)), nrow=1)

wldCount

[,1] [,2] [,3]

[1,] 174 82 85

chisq.test(wldCount)

Chi-squared test for given probabilities

data: wldCount

X-squared = 48.0762, df = 2, p-value = 3.634e-11

# Function used to plot chisq density with df=2.

dwdl <- function(x) {dchisq(x, df=2)} #dwdl is chisq distribution of win-lose-draw

# Home team summary results

aggregate(x = cbind(HomeTeamScore, VisitorScore), list(HomeTeam), FUN=sum)

# boxplot of home goal differential by team.

homeGoalDiff = mls$HomeGoalDiff

homeTeam = factor(mls$HomeAbbr)

lmfit = (homeGoalDiff ~ homeTeam)

boxplot(lmfit, main="Home team goal differential", xlab="Team", ylab="Home team score - visitor team score")

# boxplot of home goal differential as a function of home time zone - visitor time zone.

lmfit = (mls$HomeGoalDiff ~ mls$VisitorTimeZoneDiff)

boxplot(lmfit, notch=TRUE, xlab="Home time zone - Visitor time zone", ylab="Home team goal differential", main="Effect of time zone difference on home team goal differential")

# boxplot of home point differential as a function of home time zone - visitor time zone.

lmfit = (mls$HomePoints ~ mls$VisitorTimeZoneDiff)

boxplot(lmfit)

lmfit = (mls$HomePoints ~ mls$VisitorTimeZoneDiff)

boxplot(lmfit)

# Get the home goal differential and visitor time zone differential by home team.

goalDiffByTimeZoneDiff = cbind(mls$HomeTeam, mls$HomeGoalDiff, mls$VisitorTimeZoneDiff)

names(goalDiffByTimeZoneDiff) = c("HomeTeam", "HomeTeamGoalDifferential", "VisitorTimeZoneDiff")

# Get the home goal differential and visitor time zone differential by home team.

homeTeam = mls$HomeTeam

homeTeam = levels(mls$HomeTeam)

homeGoalDiff = mls$HomeGoalDiff

visitorTimeZoneDiff = mls$VisitorTimeZoneDiff

goalDiffByTimeZoneDiff = cbind(homeTeam, homeGoalDiff, visitorTimeZoneDiff)

# Home goal differential as a function of visitor time zone difference.

homeTeam = levels(mls$HomeTeam)

homeGoalDiff = mls$HomeGoalDiff

visitorTimeZoneDiff = mls$VisitorTimeZoneDiff

lmfit = lm(homeGoalDiff ~ visitorTimeZoneDiff)

plot(visitorTimeZoneDiff, jitter(homeGoalDiff), xlab="Home team time zone - Visitor team time zone", ylab="Home team goal differential", main="Home team goal differential versus visitor team time zone difference")

abline(lmfit)

goalDiffByTimeZoneDiff = cbind(levels(mls$HomeTeam), mls$HomeGoalDiff, mls$VisitorTimeZoneDiff)

names(goalDiffByTimeZoneDiff) = c("HomeTeam", "HomeTeamGoalDifferential", "VisitorTimeZoneDiff")

# Scatter plot of home goal differential as a function of visitor time zone difference.

homeTeam = levels(mls$HomeTeam)

homeGoalDiff = mls$HomeGoalDiff

visitorTimeZoneDiff = mls$VisitorTimeZoneDiff

lmfit = lm(homeGoalDiff ~ visitorTimeZoneDiff)

plot(visitorTimeZoneDiff, jitter(homeGoalDiff), xlab="Home team time zone - Visitor team time zone", ylab="Home team goal differential", main="Home team goal differential versus visitor team time zone difference")

cols = c("red", "orange", "yellow", "green", "cyan", "blue", "violet")

abline(h=mean(mls$HomeGoalDiff[which(mls$VisitorTimeZoneDiff == -3)]), col=cols[1])

abline(h=mean(mls$HomeGoalDiff[which(mls$VisitorTimeZoneDiff == -2)]), col=cols[2)

abline(h=mean(mls$HomeGoalDiff[which(mls$VisitorTimeZoneDiff == -1)]), col=cols[3)

abline(h=mean(mls$HomeGoalDiff[which(mls$VisitorTimeZoneDiff == 0)]), col=cols[4])

abline(h=mean(mls$HomeGoalDiff[which(mls$VisitorTimeZoneDiff == 1)]), col=cols[5])

abline(h=mean(mls$HomeGoalDiff[which(mls$VisitorTimeZoneDiff == 2)]), col=cols[6])

abline(h=mean(mls$HomeGoalDiff[which(mls$VisitorTimeZoneDiff == 3)]), col=cols[7])

legend("topleft", legend=c(-3:3), col=cols, pch=19, pt.cex=2, horiz=TRUE)

# Time zone difference effect sizes

summary.lm(aov(mls$HomeGoalDiff ~ mls$VisitorTimeZoneDiff))

# Effect of visitor distance travelled upon home goal differential

plot(mls$NauticalMiles, mls$HomeGoalDiff, main="Home team goal differential versus visitor travel distance",

xlab="Nautical miles traveled by visitor", ylab="Home goal differential")

abline(lm(mls$HomeGoalDiff ~ mls$NauticalMiles))

# Travel distance effect sizes

summary.lm(aov(mls$HomeGoalDiff ~ mls$NauticalMiles))

# Effect of upon average home stadium attendance upon home goal differential

plot(mls$HomeTeamAttendance, mls$HomeGoalDiff, main="Home team goal differential versus home team attendance",

xlab="Average home team attendance", ylab="Home goal differential")

abline(lm(mls$HomeGoalDiff ~ mls$HomeTeamAttendance))

# Average attendance effect size

summary.lm(aov(mls$HomeGoalDiff ~ mls$HomeTeamAttendance))

# Plot the almost significant attendance

dfx = function(x) {df(x, 1, 339)}

curve(dfx, from=0, to=5)

abline(v=3.346)

# Effect of upon average home stadium capacity upon home goal differential

plot(mls$HomeTeamCapacity, mls$HomeGoalDiff, main="Home team goal differential versus home team stadium capacity",

xlab="Average home team stadium capacity", ylab="Home goal differential")

abline(lm(mls$HomeGoalDiff ~ mls$HomeTeamCapacity))

# Stadium capacity effect size

summary.lm(aov(mls$HomeGoalDiff ~ mls$HomeTeamCapacity))

# Effect of upon average home stadium utilization upon home goal differential

plot(mls$FractionFull, mls$HomeGoalDiff, main="Home team goal differential versus home team stadium utilization",

xlab="Average home team stadium utilization", ylab="Home goal differential")

abline(lm(mls$HomeGoalDiff ~ mls$FractionFull))

# Average stadium utilization effect size

summary.lm(aov(mls$HomeGoalDiff ~ mls$FractionFull))

# Effect of home team salary upon home goal differential

plot(mls$HomeTeamSalary, mls$HomeGoalDiff, main="Home team goal differential versus home team salary",

xlab="Team Salary", ylab="Home goal differential")

abline(lm(mls$HomeGoalDiff ~ mls$HomeTeamSalary))

# Average home team salary effect size

summary.lm(aov(mls$HomeGoalDiff ~ mls$HomeTeamSalary))

# Appendix B: Screen Scraper Script for [www.goal.com](http://www.goal.com)

This script was written in Windows PowerShell to fix content copied from [www.goal.com](http://www.goal.com). It takes content from the web site, and transforms it into a CSV format that can be consumed by R and Excel.

To use the script:

Load a typical page into a web browser:

http://www.goal.com/en-us/teams/united-states/1364/seattle-sounders-fc/calendar

Copy all the match results, and paste them into a text file.

Run the script below with the following arguments:

Path to the text file you just created.

Path to the CSV text file that you want produced.

Example: datafix.ps1 soundersFC.txt soundersFC.csv

#===datafix.ps1=============================================

param (

[string] $srcPath,

[string] $destPath

)

[string[]] $lines = Get-Content $srcPath

# Raw content looks like this (minus the # symbols):

#Mar 10, 2012 USA - Major League Soccer

#Completed Colorado Rapids2 - 0 Columbus Crew

#Mar 24, 2012 USA - Major League Soccer

#Completed Columbus Crew2 - 0 Montreal Impact

# Join consecutive pairs of lines into a single line

$jLines = @();

for ($i = 0; $i -lt $lines.Count; $i += 2) {$jlines += $lines[$i] + ' ' + $lines[$i + 1] }

# Each jLine looks like this:

#Mar 10, 2012 USA - Major League Soccer Completed Colorado Rapids2 - 0 Columbus Crew

# Fix the various problems to make the content CSV compatible.

[string[]] $csvLines = @()

[string] $line

foreach ($line in $jLines) {

# Fix the month part of the time stamp. We want this format: 3/17/2012.

if ($line -match "^Mar\s+") {$line = $line -replace "^Mar\s+", "3/"}

elseif ($line -match "^Apr\s+") {$line = $line -replace "^Apr\s+", "4/"}

elseif ($line -match "^May\s+") {$line = $line -replace "^May\s+", "5/"}

elseif ($line -match "^Jun\s+") {$line = $line -replace "^Jun\s+", "6/"}

elseif ($line -match "^Jul\s+") {$line = $line -replace "^Jul\s+", "7/"}

elseif ($line -match "^Aug\s+") {$line = $line -replace "^Aug\s+", "8/"}

elseif ($line -match "^Sep\s+") {$line = $line -replace "^Sep\s+", "9/"}

elseif ($line -match "^Oct\s+") {$line = $line -replace "^Oct\s+", "10/"}

elseif ($line -match "^Nov\s+") {$line = $line -replace "^Nov\s+", "11/"}

elseif ($line -match "^(\d+)-(\d+)-(\d+)\s+(.+)") {$line = $line -replace "^(\d+)-(\d+)-(\d+)\s+(.+)", '$2/$3/$1,$4'}

# $line now looks like this:

#3/10, 2012 USA - Major League Soccer Completed Colorado Rapids2 - 0 Columbus Crew

# Fix the year, and get rid of everything through "Completed ".

$line = [regex]::Replace($line, ",\s+2012.+Completed\s+", "/2012,")

$line = [regex]::Replace($line, "Major\s+League\s+Soccer\s+Completed\s+", "")

# $line now looks like this:

#3/10/2012,Colorado Rapids2 - 0 Columbus Crew

# Put a space between the team name and the score.

$line = [regex]::Replace($line, "([a-zA-Z])(\d+)", '$1 $2');

# $line now looks like this:

#3/10/2012,Colorado Rapids 2 - 0 Columbus Crew

# Put a comma after the home score and visitor score

$line = [regex]::Replace($line, "\s+(\d+)\s+-\s+(\d+)\s", ',$1,$2,');

# $line now looks like this:

#3/10/2012,Colorado Rapids,2,0,Columbus Crew

# Good to go into the CSV array.

$csvLines += $line

}

# Write the $csvLines array to the destination.

#$csvLines

$csvLines | Set-Content $destPath