Applied Statistics and NHL Data

Alex Diaz-Papkovich
Carleton University
alexdiaz.ca
@adp_diaz

Introduction

Introduction

- Universal truths:

Introduction

- Universal truths:
- Math is difficult

Introduction

- Universal truths:
  - Math is difficult
  - Hockey is fun

Introduction

- Universal truths:
  - Math is difficult
  - Hockey is fun
- My motivation: Using NHL data to study statistics

NHL data as a learning tool

NHL data as a learning tool

- Toyed with it in undergrad

NHL data as a learning tool

- Toyed with it in undergrad
- Same datasets often get used - beetles, flower petals, common public datasets

NHL data as a learning tool

- Toyed with it in undergrad
- Same datasets often get used - beetles, flower petals, common public datasets
- Using a dataset you enjoy enhances what you learn

NHL data as a learning tool

- Toyed with it in undergrad
- Same datasets often get used - beetles, flower petals, common public datasets
- Using a dataset you enjoy enhances what you learn
    - Provides understandable, concrete connections

The process

The process

- Stages of statistical analysis for a grad student?

The process

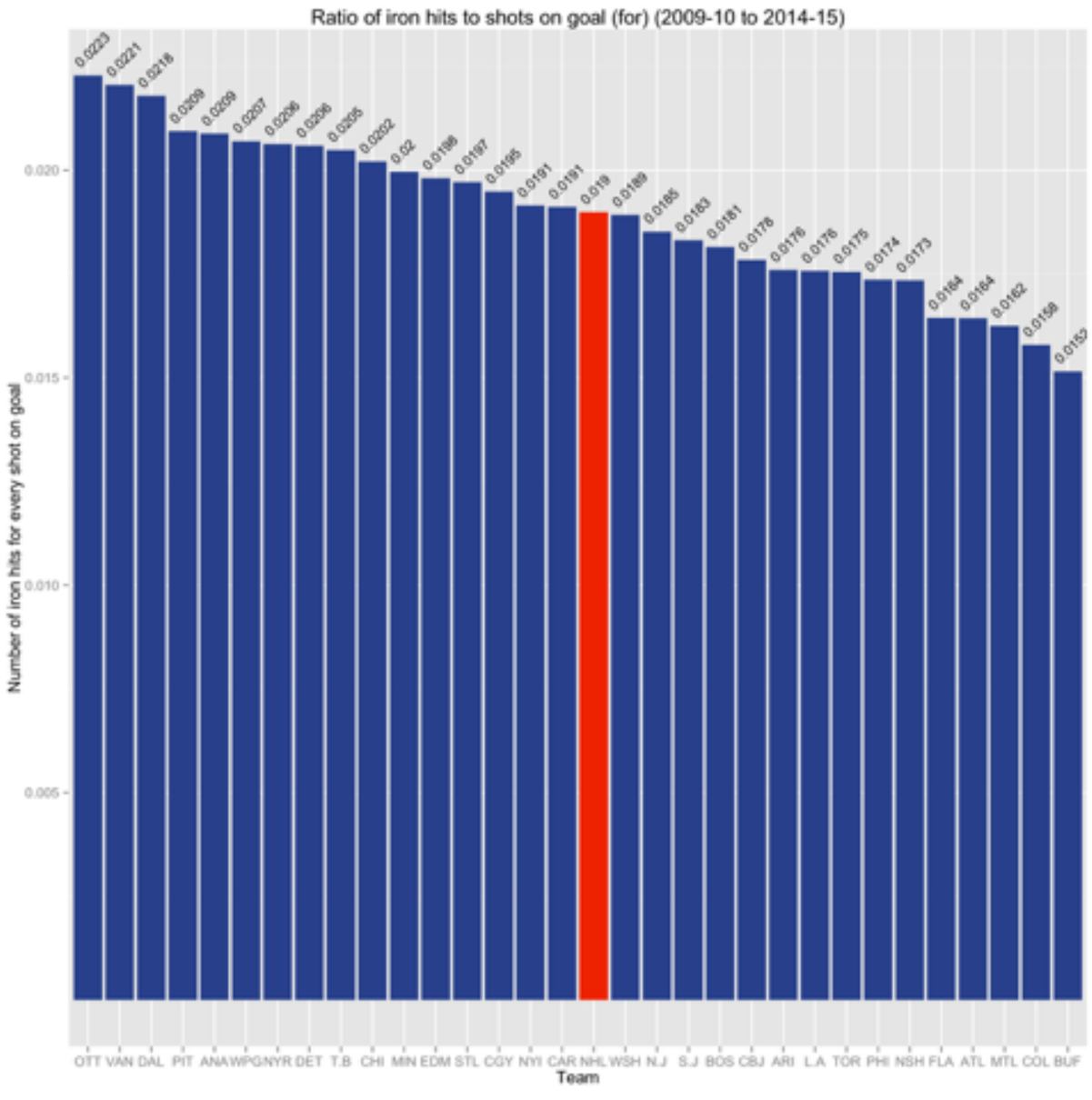
- Stages of statistical analysis for a grad student?
- Denial, anger, bargaining, depression, acceptance?

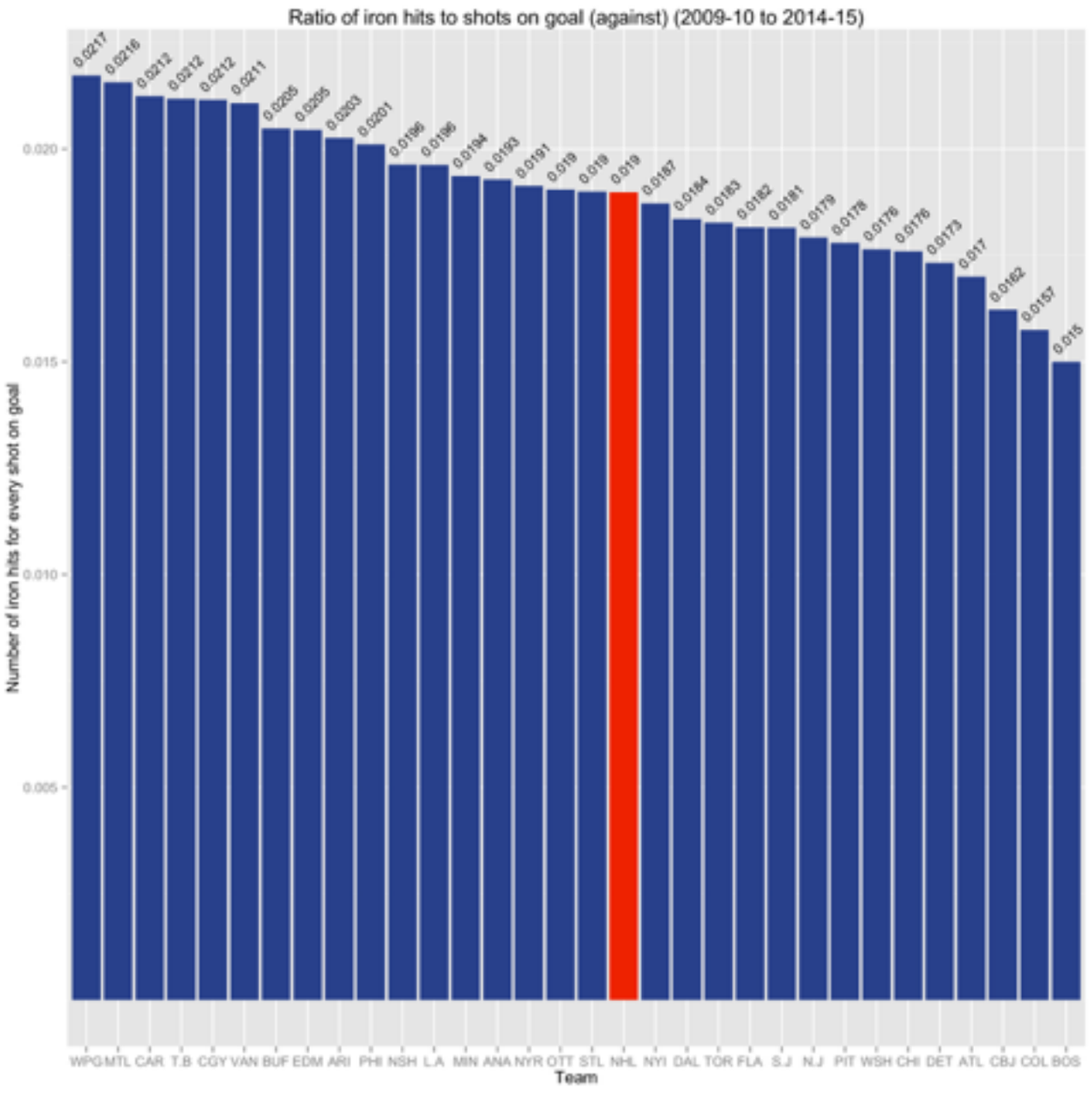
The process

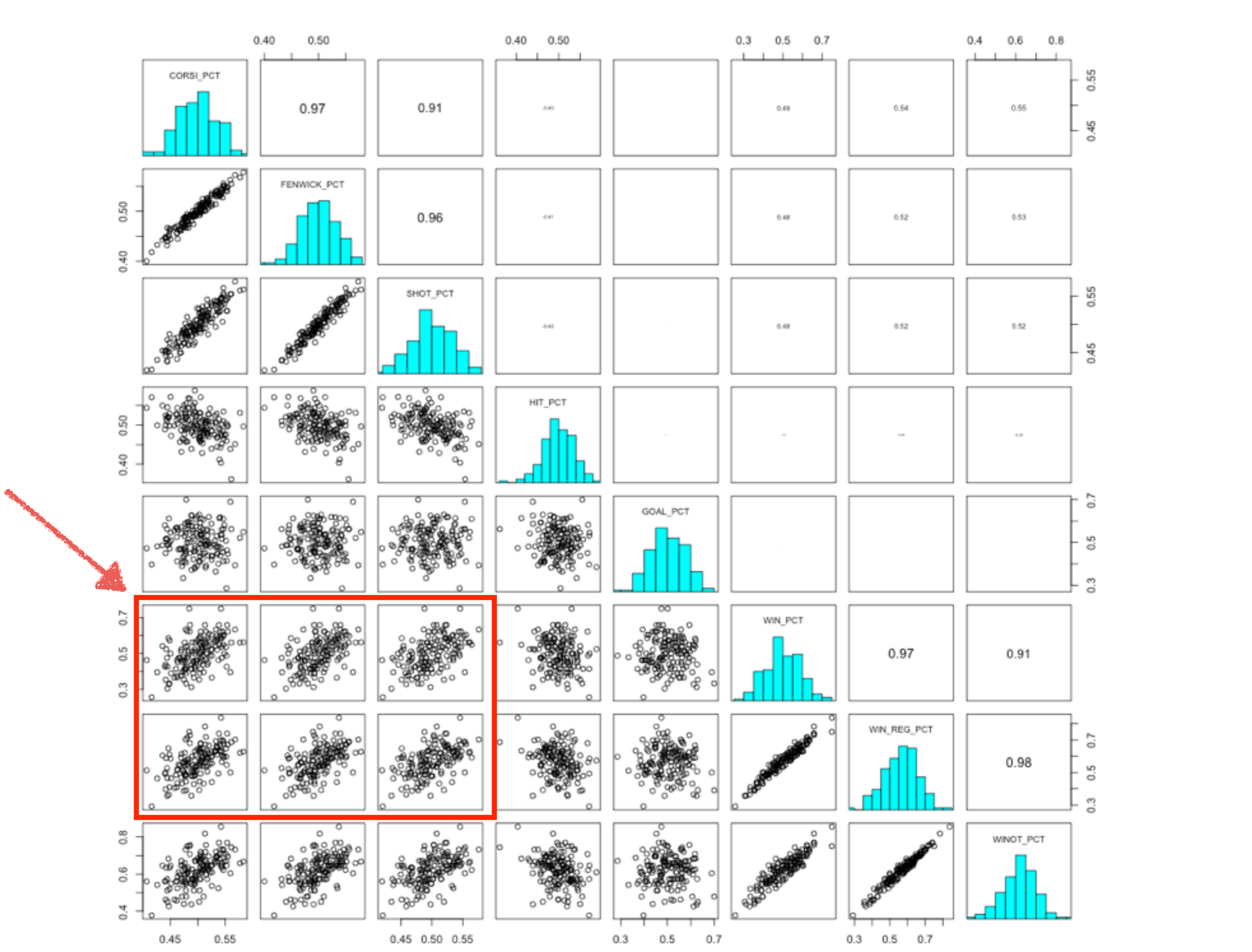
- Stages of statistical analysis for a grad student?
- Denial, anger, bargaining, depression, acceptance?

- Broadly: data preparation, exploration, modelling, analysis, evaluation

Who pays the iron price?






CORSI_PCT
FENWICK_PCT
SHOT_PCT
HIT_PCT
GOAL_PCT
WIN_PCT
WIN_REG_PCT
WINOT_PCT
0.97
0.91
0.96
0.97
0.91
0.98

Data preparation

Data preparation

- Often 80%+ of the work

Data preparation

- Often 80%+ of the work
- Stages of data preparation:

Data preparation

- Often 80%+ of the work
- Stages of data preparation:
  - Anger, anger, anger, anger, acceptance

Data preparation

- Often 80%+ of the work
- Stages of data preparation:
  - Anger, anger, anger, anger, acceptance
    - “What do you mean the RTSS doesn’t track which team went offside?”

Data preparation

- Often 80%+ of the work
- Stages of data preparation:
  - Anger, anger, anger, anger, acceptance
    - “What do you mean the RTSS doesn’t track which team went offside?”
    - “There are blank player names??”

Data preparation

- Often 80%+ of the work
- Stages of data preparation:
  - Anger, anger, anger, anger, acceptance
    - “What do you mean the RTSS doesn’t track which team went offside?”
    - “There are blank player names??”
    - “Why does the 2nd period end at -16:0-1???”

Modelling

Modelling

- Make use of common techniques

Modelling

- Make use of common techniques
  - Linear regression

Modelling

- Make use of common techniques
  - Linear regression
  - Logistic regression

Modelling

- Make use of common techniques
  - Linear regression
  - Logistic regression
  - Hypothesis testing

Modelling

- Make use of common techniques
  - Linear regression
  - Logistic regression
  - Hypothesis testing
- Good idea to challenge some assumptions

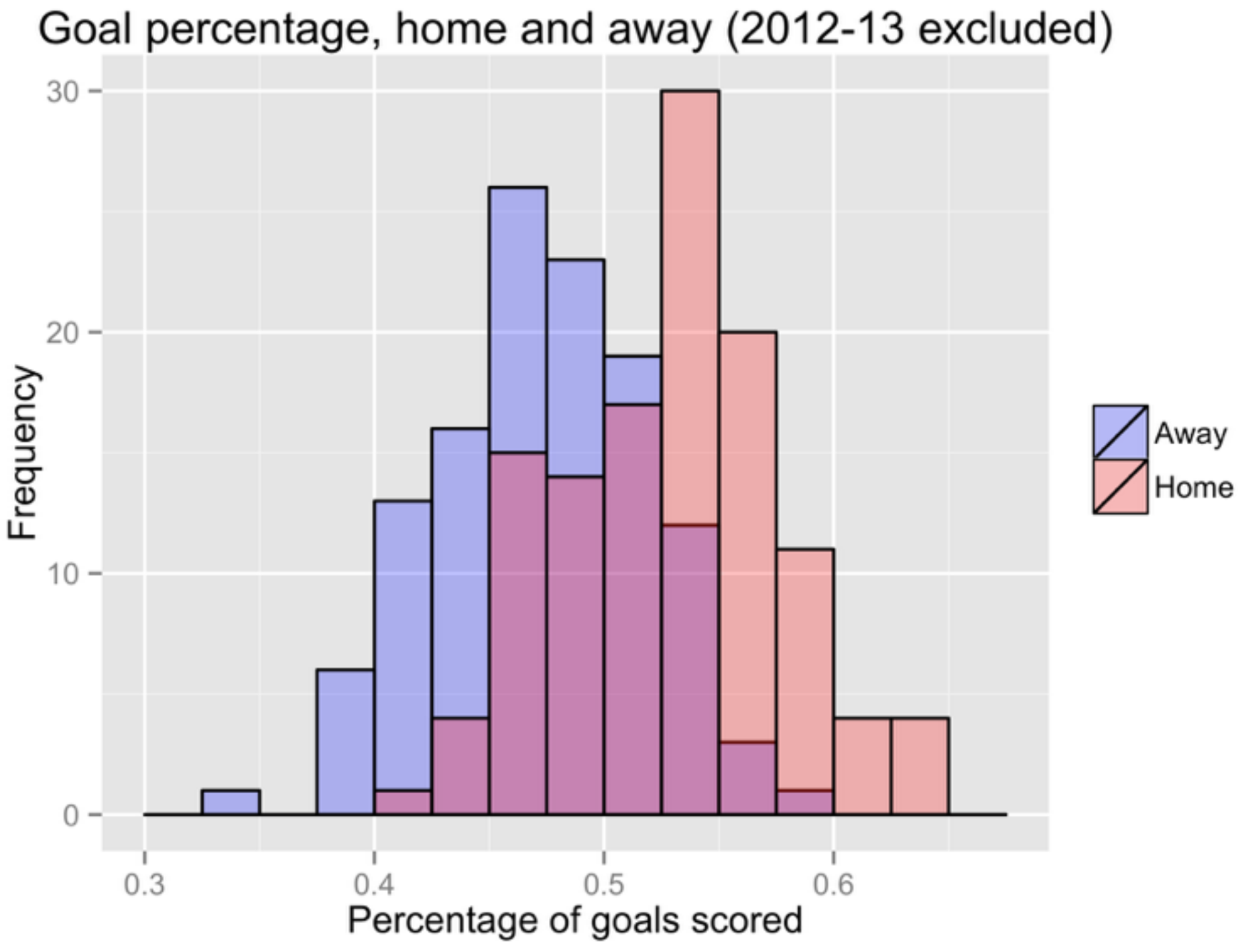
Modelling

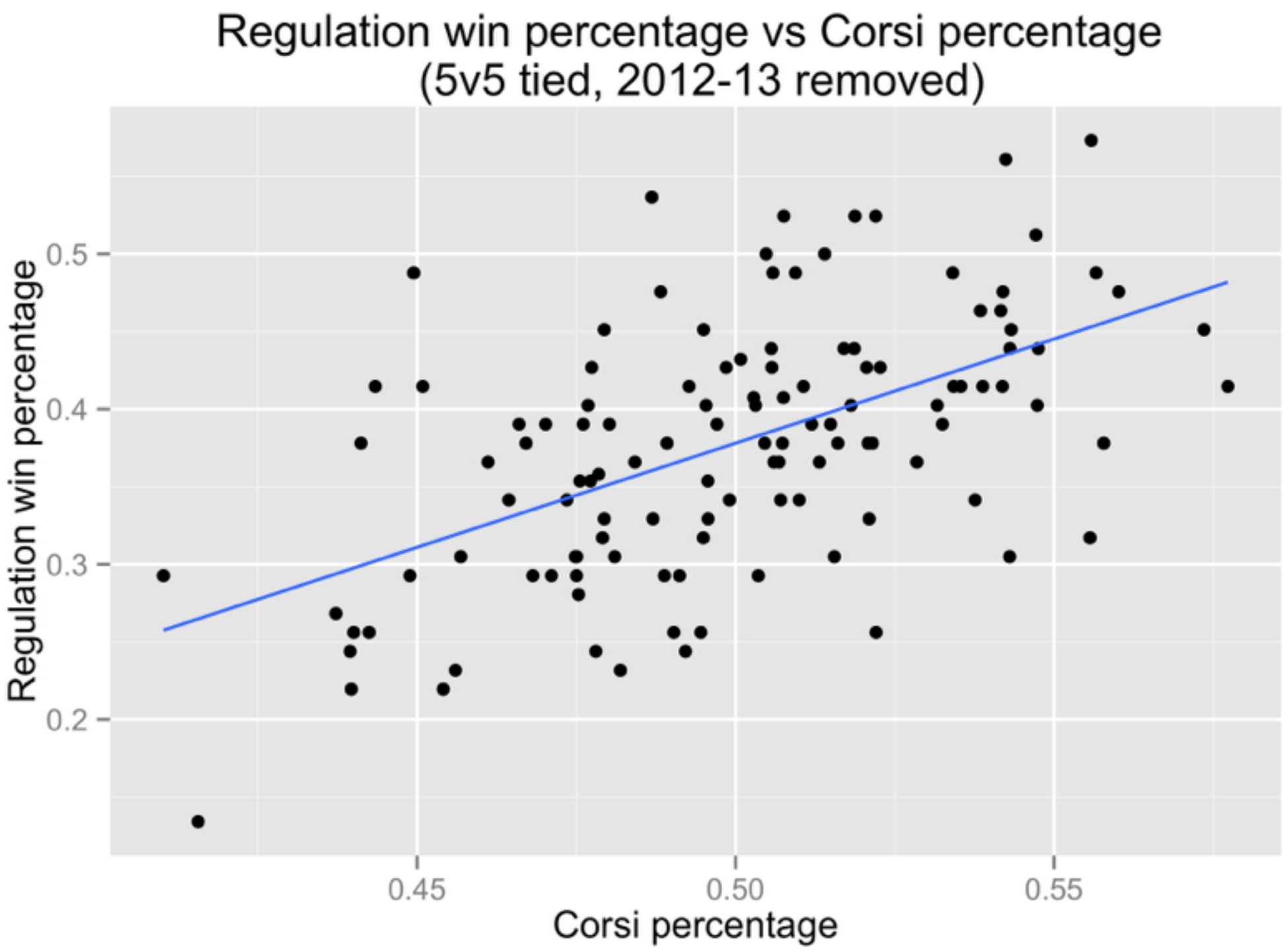
- Make use of common techniques
  - Linear regression
  - Logistic regression
  - Hypothesis testing
- Good idea to challenge some assumptions
  - Is Corsi *actually* a useful measure?

Modelling

- Make use of common techniques
  - Linear regression
  - Logistic regression
  - Hypothesis testing
- Good idea to challenge some assumptions
  - Is Corsi *actually* a useful measure?
  - Does home ice advantage exist?

Home ice


Goal percentage, home and away (2012-13 excluded)
Frequency
30
20
10
0
0.3
0.4
0.5
0.6
Percentage of goals scored
Away
Home


Regulation win percentage vs Corsi percentage
(5v5 tied, 2012-13 removed)
Regulation win percentage
0.5
0.4
0.3
0.2
0.45
0.50
0.55
Corsi percentage

Comparing shot metrics

	Dependent variable:		
	Win%		
	(1)	(2)	(3)
Shot%	1.347***		
(Std. Err.)	(0.193)		
Fenwick%		1.304***	
(Std. Err.)		(0.192)	
Corsi%			1.342***
(Std. Err.)			(0.187)
Constant	−0.296***	−0.274***	−0.293***
(Std. Err.)	(0.097)	(0.096)	(0.094)
Observations	120	120	120
R^2	0.291	0.282	0.303
Adjusted R^2	0.285	0.276	0.297
Residual Std. Error (df = 118)	0.070	0.070	0.069
F Statistic (df = 1; 118)	48.533***	46.293***	51.372***
Note:			*p<0.1; **p<0.05; ***p<0.01
			(1), (2), and (3) are model IDs for each metric

Other factors: PP vs PK

	Dependent variable:			
	Win%			
	(1)	(2)	(3)	(4)
PP goals for	0.908***		0.885***	1.978*
(Std. Err.)	(0.193)		(0.177)	(1.012)
PK goals against		−0.807***	−0.785***	0.232
(Std. Err.)		(0.176)	(0.161)	(0.941)
PP/PK interaction				−4.897
(Std. Err.)				(4.463)
Constant	0.181***	0.554***	0.358***	0.130
(Std. Err.)	(0.042)	(0.039)	(0.053)	(0.214)
Observations	120	120	120	120
R^2	0.157	0.151	0.300	0.307
Adjusted R^2	0.150	0.144	0.288	0.289
Residual Std. Error	0.076	0.076	0.070	0.070
Residual Std. Error DF	118	118	117	116
F Statistic	22.025***	20.956***	25.084***	17.153***
F Statistic DF	(1;118)	(1;118)	(2;117)	(3; 116)

Note: *p<0.1; **p<0.05; ***p<0.01

(1), (2), (3), and (4) are model IDs for special teams and interactions

Thinking about theory

Thinking about theory

- What does a shot attempt represent? Why is it useful to measure?

Thinking about theory

- What does a shot attempt represent? Why is it useful to measure?
- Used as a proxy for possession, but what else does it mean?

Thinking about theory

- What does a shot attempt represent? Why is it useful to measure?
    - Used as a proxy for possession, but what else does it mean?
        - Control the game, forcing opposition to react

Thinking about theory

- What does a shot attempt represent? Why is it useful to measure?
    - Used as a proxy for possession, but what else does it mean?
        - Control the game, forcing opposition to react
- Ultimately, justify decisions in further analysis

Association rule learning

Association rule learning

- Data mining technique that works on large binary databases

Association rule learning

- Data mining technique that works on large binary databases
- Previewed at last year's conference

Association rule learning

- Data mining technique that works on large binary databases
- Previewed at last year's conference
- Think of it as a large-scale With Or Without You (WOWY)

Association rule learning

- Data mining technique that works on large binary databases
- Previewed at last year's conference
- Think of it as a large-scale With Or Without You (WOWY)
- SUPP(X) = % of transactions with itemset X

Association rule learning

- Data mining technique that works on large binary databases
- Previewed at last year's conference
- Think of it as a large-scale With Or Without You (WOWY)
- SUPP(X) = % of transactions with itemset X

  = Pr(X) = Probability X happens

Association rule learning

- Data mining technique that works on large binary databases
- Previewed at last year's conference
- Think of it as a large-scale With Or Without You (WOWY)
- SUPP(X) = % of transactions with itemset X

  = Pr(X) = Probability X happens
- LIFT(X,Y) = SUPP(X,Y)/[SUPP(X)*SUPP(Y)]

Association rule learning

- Data mining technique that works on large binary databases
- Previewed at last year's conference
- Think of it as a large-scale With Or Without You (WOWY)
- SUPP(X) = % of transactions with itemset X

  = Pr(X) = Probability X happens
- LIFT(X,Y) = SUPP(X,Y)/[SUPP(X)*SUPP(Y)]

  = Measure of independence (farther from 1 => less independent)

Association rule learning

- Data mining technique that works on large binary databases
- Previewed at last year's conference
- Think of it as a large-scale With Or Without You (WOWY)
- SUPP(X) = % of transactions with itemset X

  = Pr(X) = Probability X happens
- LIFT(X,Y) = SUPP(X,Y)/[SUPP(X)*SUPP(Y)]

  = Measure of independence (farther from 1 => less independent)
- Events {A,B} are independent if Pr(A and B) = Pr(A) x Pr(B)

Example database

COLTON_ORR	JOHN_MICHAEL_LILES	JAMES_VAN_RIEMSDYK	SHOT_FOR	SHOT_AGAINST	MISS_FOR	MISS_AGAINST	BLOCK_FOR	BLOCK_AGAINST
0	0	0	0	0	1	0	0	0
0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	1	0
1	0	0	0	0	0	0	1	0
0	0	1	0	1	0	0	0	0
0	0	1	0	1	0	0	0	0
0	0	1	1	0	0	0	0	0

Example database

COLTON_ORR	JOHN_MICHAEL_LILES	JAMES_VAN_RIEMSDYK	SHOT_FOR	SHOT_AGAINST	MISS_FOR	MISS_AGAINST	BLOCK_FOR	BLOCK_AGAINST
0	0	0	0	0	1	0	0	0
0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	1	0
1	0	0	0	0	0	0	1	0
0	0	1	0	1	0	0	0	0
0	0	1	0	1	0	0	0	0
0	0	1	1	0	0	0	0	0

- Created a database of “transactions”

Example database

COLTON_ORR	JOHN_MICHAEL_LILES	JAMES_VAN_RIEMSDYK	SHOT_FOR	SHOT_AGAINST	MISS_FOR	MISS_AGAINST	BLOCK_FOR	BLOCK_AGAINST
0	0	0	0	0	1	0	0	0
0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	1	0
1	0	0	0	0	0	0	1	0
0	0	1	0	1	0	0	0	0
0	0	1	0	1	0	0	0	0
0	0	1	1	0	0	0	0	0

- Created a database of “transactions”
- Players on the left, events on the right

Example database

COLTON_ORR	JOHN_MICHAEL_LILES	JAMES_VAN_RIEMSDYK	SHOT_FOR	SHOT_AGAINST	MISS_FOR	MISS_AGAINST	BLOCK_FOR	BLOCK_AGAINST
0	0	0	0	0	1	0	0	0
0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	1	0
1	0	0	0	0	0	0	1	0
0	0	1	0	1	0	0	0	0
0	0	1	0	1	0	0	0	0
0	0	1	1	0	0	0	0	0

- Created a database of “transactions”
- Players on the left, events on the right
- Looking at Corsi events (EV, 5v5, tied)

Example database

COLTON_ORR	JOHN_MICHAEL_LILES	JAMES_VAN_RIEMSDYK	SHOT_FOR	SHOT_AGAINST	MISS_FOR	MISS_AGAINST	BLOCK_FOR	BLOCK_AGAINST
0	0	0	0	0	1	0	0	0
0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	1	0
1	0	0	0	0	0	0	1	0
0	0	1	0	1	0	0	0	0
0	0	1	0	1	0	0	0	0
0	0	1	1	0	0	0	0	0

- Created a database of “transactions”
- Players on the left, events on the right
- Looking at Corsi events (EV, 5v5, tied)
- Basically try to figure out what players drive the play within a line, team, defense pairing, etc

2013-14 Toronto Maple Leafs probability of events, by all players

Corsi (against)

Corsi (for)

Player	Corsi (against)	Corsi (for)
PHIL_KESSEL	20%	15.9%
DION_PHANEUF	20.	13.6%
JAMES_VAN_RIEMSDYK	18.9%	15.2%
JAKE_GARDINER	19.5%	13.7%
CARL_GUNNARSSON	18.7%	12.9%
NAZEM_KADRI	17.2%	12.9%
CODY_FRANSON	17.1%	12.6%
MASON_RAYMOND	15.9%	12%
MORGAN_RIELLY	16.2%	11.4%
JOFFREY_LUPUL	16.3%	9.6%
TYLER_BOZAK	14.5%	11.2%
DAVID_CLARKSON	13.2%	8.9%
NIKOLAI_KULEMIN	13.4%	8.6%
JAY_MCCLEMENT	13.5%	7.2%
PAUL_RANGER	10.4%	7.5%
TIM_GLEASON	9.4%	6.5%
PETER_HOLLAND	6.2%	4.2%
TROY_BODIE	5.2%	3.8%
DAVE_BOLLAND	4.1%	3.1%
MARK_FRASER	4.2%	2.8%
COLTON_ORR	3.8%	2.1%
CARTER_ASHTON	2.9%	1.9%
TREVOR_SMITH	2.8%	1.7%
JERRY_DAMIGO	2%	1.5%
FRAZER_MCLAREN	1.4%	
JERRED_SMITHSON	1.3%	
JOHN_MICHAEL_LILES	1.1%	
DAVID_BROLL	0.9%	

Lift

1.1

1.0

0.9

alexdiaz.ca
@adp_diaz

0.00 0.05 0.10 0.15 0.20

Support for {PLAYER, EVENT} combination

2013-14 Toronto Maple Leafs association rules, by forward lines

PHIL_KESSEL,NAZEM_KADRI,JAMES_VAN_RIEMSDYK

Corsi (for)

Per 60: 64.49
(Lift: 1.271)

← Corsi for

Corsi (against)

Per 60: 59.18
(Lift: 0.811)

← Corsi against

Lift
1.3
1.2
1.1
1.0
0.9
0.8

Players

TYLER_BOZAK,PHIL_KESSEL,JAMES_VAN_RIEMSDYK

Corsi (for)

Per 60: 56.6
(Lift: 1.136)

Corsi (against)

Per 60: 64.91
(Lift: 0.906)

0.00 0.03 0.06 0.09 0.12

Support for {PLAYER, EVENT} combination

2014-15 Toronto Maple Leafs association rules, by forward lines

JOFFREY_LUPUL,DANIEL_WINNIK,NAZEM_KADRI

Corsi (for)

Per 60: 90.36
(Lift: 1.46)

Corsi (against)

Per 60: 45.97
(Lift: 0.617)

MIKE_SANTORELLI,DANIEL_WINNIK,NAZEM_KADRI

Corsi (for)

Per 60: 81.12
(Lift: 1.305)

Corsi (against)

Per 60: 55.86
(Lift: 0.747)

PHIL_KESSEL,TYLER_BOZAK,JAMES_VAN_RIEMSDYK

Corsi (for)

Per 60: 55.59
(Lift: 0.916)

Corsi (against)

Per 60: 78.16
(Lift: 1.07)

Players

Lift
1.50
1.25
1.00
0.75

0.00
0.05
0.10

Support for {PLAYER, EVENT} combination

2014-15 Toronto Maple Leafs association rules, by forward pairs (without Kadri)

PHIL_KESSEL,JAMES_VAN_RIEMSDYK

Corsi (for)

Per 60: 56.11
(Lift: 0.928)

Corsi (against)

Per 60: 77.15
(Lift: 1.06)

PHIL_KESSEL,TYLER_BOZAK

Corsi (for)

Per 60: 52.66
(Lift: 0.902)

Corsi (against)

Per 60: 75.91
(Lift: 1.081)

TYLER_BOZAK,JAMES_VAN_RIEMSDYK

Corsi (for)

Per 60: 53.83
(Lift: 0.921)

Corsi (against)

Per 60: 74.98
(Lift: 1.066)

Players

Lift
1.2
1.0
0.8

0.00
0.05
0.10
0.15

Support for {PLAYER, EVENT} combination

2014-15 Toronto Maple Leafs association rules, by forward pairs (with Kadri)

DANIEL_WINNIK,NAZEM_KADRI

Corsi (for)

Per 60: 69.84
(Lift: 1.244)

Corsi (against)

Per 60: 53.89
(Lift: 0.798)

JOFFREY_LUPUL,NAZEM_KADRI

Corsi (for)

Per 60: 75.15
(Lift: 1.306)

Corsi (against)

Per 60: 51.58
(Lift: 0.745)

MIKE_SANTORELLI,NAZEM_KADRI

Corsi (for)

Per 60: 65.5
(Lift: 1.174)

Corsi (against)

Per 60: 57.38
(Lift: 0.855)

Players

Lift
1.2
1.0
0.8

0.00
0.02
0.04
0.06

Support for {PLAYER, EVENT} combination

2014-15 Toronto Maple Leafs association rules, by individual forwards

JAMES_VAN_RIEMSDYK

Corsi (for)

Per 60: 52.91
(Lift: 0.927)

Corsi (against)

Per 60: 72.83
(Lift: 1.061)

NAZEM_KADRI

Corsi (for)

Per 60: 60.91
(Lift: 1.169)

Corsi (against)

Per 60: 53.89
(Lift: 0.86)

PHIL_KESSEL

Corsi (for)

Per 60: 53.29
(Lift: 0.941)

Corsi (against)

Per 60: 71.45
(Lift: 1.049)

TYLER_BOZAK

Corsi (for)

Per 60: 51.56
(Lift: 0.918)

Corsi (against)

Per 60: 72.18
(Lift: 1.068)

Lift
1.2
1.1
1.0
0.9
0.8

Player

0.00
0.05
0.10
0.15
0.20

Support for {PLAYER, EVENT} combination

2015-16 Chicago Blackhawks association rules, by forwards lines

ANDREW_SHAW,MARCUS_KRUGER,ANDREW_DESJARDINS

Corsi (for) – Per 60: 45.22 (Lift: 0.952)

Corsi (against) – Per 60: 46.59 (Lift: 1.052)

ANDREW_SHAW,MARIAN_HOSSA,JONATHAN_TOEWS

Corsi (for) – Per 60: 57.21 (Lift: 1.13)

Corsi (against) – Per 60: 40.57 (Lift: 0.86)

ARTEMI_PANARIN,PATRICK_KANE,ARTEM_ANISIMOV

Corsi (for) – Per 60: 60.75 (Lift: 1.116)

Corsi (against) – Per 60: 44.37 (Lift: 0.875)

TEUVO_TERAVAINEN,MARIAN_HOSSA,JONATHAN_TOEWS

Corsi (for) – Per 60: 61.39 (Lift: 0.996)

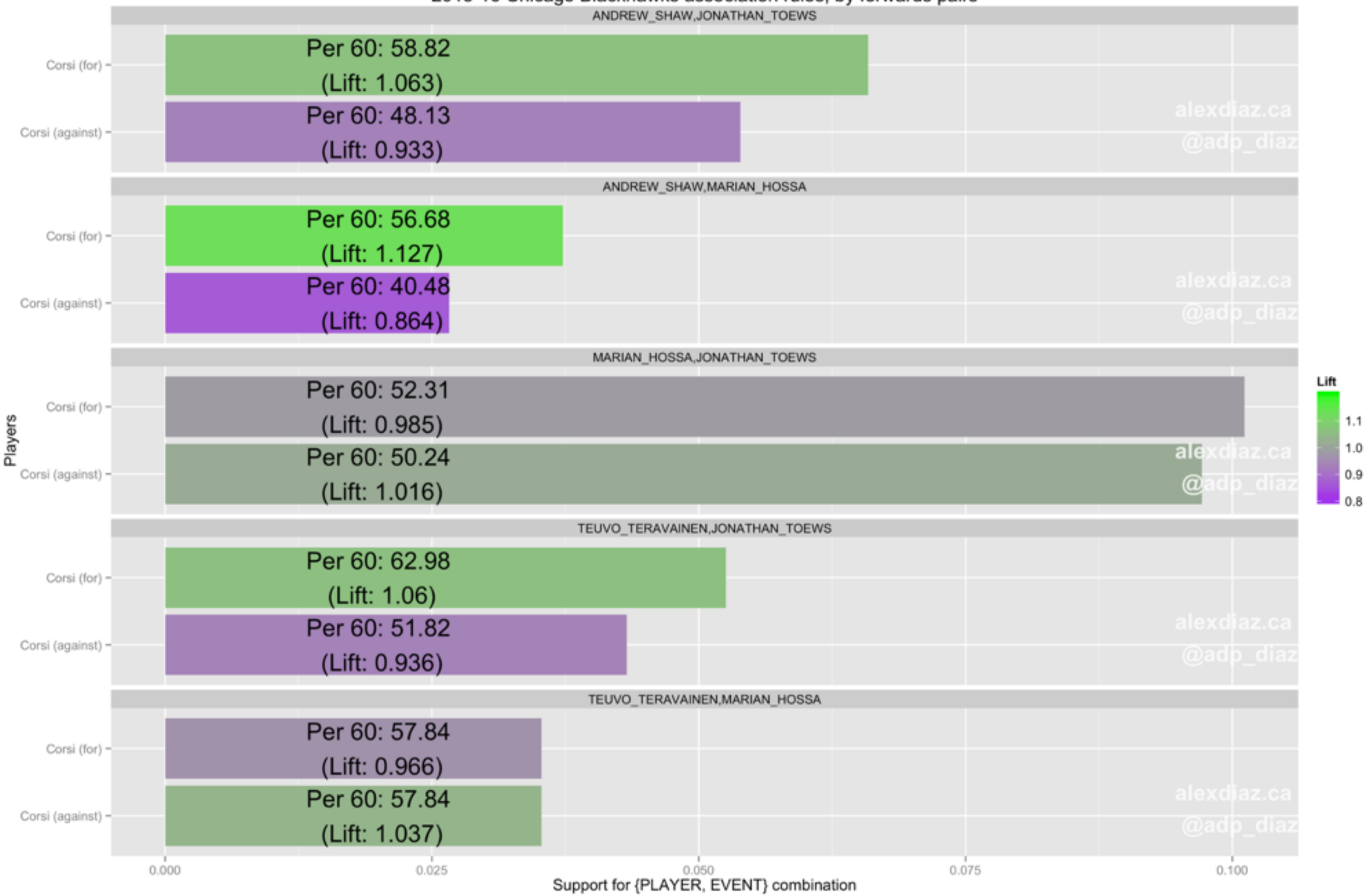
Corsi (against) – Per 60: 57.71 (Lift: 1.004)

Lift: 1.2, 1.1, 1.0, 0.9, 0.8

Player

0.00 0.05 0.10

Support for {PLAYER, EVENT} combination


2015-16 Chicago Blackhawks association rules, by forwards pairs
ANDREW_SHAW,JONATHAN_TOEWS
Corsi (for)
Per 60: 58.82
(Lift: 1.063)
Corsi (against)
Per 60: 48.13
(Lift: 0.933)
ANDREW_SHAW,MARIAN_HOSSA
Corsi (for)
Per 60: 56.68
(Lift: 1.127)
Corsi (against)
Per 60: 40.48
(Lift: 0.864)
MARIAN_HOSSA,JONATHAN_TOEWS
Corsi (for)
Per 60: 52.31
(Lift: 0.985)
Corsi (against)
Per 60: 50.24
(Lift: 1.016)
TEUVO_TERAVAINEN,JONATHAN_TOEWS
Corsi (for)
Per 60: 62.98
(Lift: 1.06)
Corsi (against)
Per 60: 51.82
(Lift: 0.936)
TEUVO_TERAVAINEN,MARIAN_HOSSA
Corsi (for)
Per 60: 57.84
(Lift: 0.966)
Corsi (against)
Per 60: 57.84
(Lift: 1.037)
alexdiaz.ca
@adp_diaz
Players
Lift
1.1
1.0
0.9
0.8
0.000
0.025
0.050
0.075
0.100
Support for {PLAYER, EVENT} combination

2015-16 Pittsburgh Penguins association rules, by forward lines

DAVID_PERRON,EVGENI_MALKIN,PHIL_KESSEL

Corsi (for)

Per 60: 63.49
(Lift: 1.049)

Corsi (against)

Per 60: 53.77
(Lift: 0.948)

DAVID_PERRON,SIDNEY_CROSBY,CHRIS_KUNITZ

Corsi (for)

Per 60: 69.17
(Lift: 1.178)

Corsi (against)

Per 60: 44.58
(Lift: 0.81)

Lift
1.2
1.1
1.0
0.9
0.8

Play

SIDNEY_CROSBY,PATRIC_HORNQVIST,PASCAL_DUPUIS

Corsi (for)

Per 60: 71.21
(Lift: 1.119)

Corsi (against)

Per 60: 52.08
(Lift: 0.873)

SIDNEY_CROSBY,PHIL_KESSEL,CHRIS_KUNITZ

Corsi (for)

Per 60: 50.51
(Lift: 0.868)

Corsi (against)

Per 60: 62.17
(Lift: 1.141)

0.00 0.02 0.04 0.06 0.08

Support for {PLAYER, EVENT} combination

2015-16 Pittsburgh Penguins association rules, by forward pairs (Crosby)

SIDNEY_CROSBY,CHRIS_KUNITZ

Corsi (for)

Per 60: 64.91
(Lift: 1.077)

Corsi (against)

Per 60: 51.81
(Lift: 0.918)

SIDNEY_CROSBY,PASCAL_DUPUIS

Corsi (for)

Per 60: 62.86
(Lift: 0.999)

Corsi (against)

Per 60: 59.03
(Lift: 1.001)

SIDNEY_CROSBY,PATRIC_HORNQVIST

Corsi (for)

Per 60: 63.65
(Lift: 1.034)

Corsi (against)

Per 60: 55.6
(Lift: 0.964)

SIDNEY_CROSBY,PHIL_KESSEL

Corsi (for)

Per 60: 50.04
(Lift: 0.788)

Corsi (against)

Per 60: 72.92
(Lift: 1.226)

Lift
1.3
1.2
1.1
1.0
0.9
0.8
0.7

Player

0.000
0.025
0.050
0.075

Support for {PLAYER, EVENT} combination

Evaluation

Evaluation

- Doesn’t account for quality of opposition or zone starts

Evaluation

- Doesn’t account for quality of opposition or zone starts
- Doesn’t directly use time-on-ice

Evaluation

- Doesn’t account for quality of opposition or zone starts
- Doesn’t directly use time-on-ice
- No clear “thresholds”

Evaluation

- Doesn’t account for quality of opposition or zone starts
- Doesn’t directly use time-on-ice
- No clear “thresholds”
    - Needs variance estimation, especially for lower sample sizes

Evaluation

- Doesn’t account for quality of opposition or zone starts
- Doesn’t directly use time-on-ice
- No clear “thresholds”
    - Needs variance estimation, especially for lower sample sizes
- Still useful and informative

Conclusions

Conclusions

- Useful for measuring player contributions within a team

Conclusions

- Useful for measuring player contributions within a team
- Potential uses for coaching, management

Conclusions

- Useful for measuring player contributions within a team
- Potential uses for coaching, management
  - Measure player performance within given context

Conclusions

- Useful for measuring player contributions within a team
- Potential uses for coaching, management
  - Measure player performance within given context
  - Create more effective lines

Conclusions

- Useful for measuring player contributions within a team
- Potential uses for coaching, management
  - Measure player performance within given context
  - Create more effective lines
  - Make decisions between players with similar roles

Questions/Comments/ Concerns?

- alexdiaz.ca

- @adp_diaz