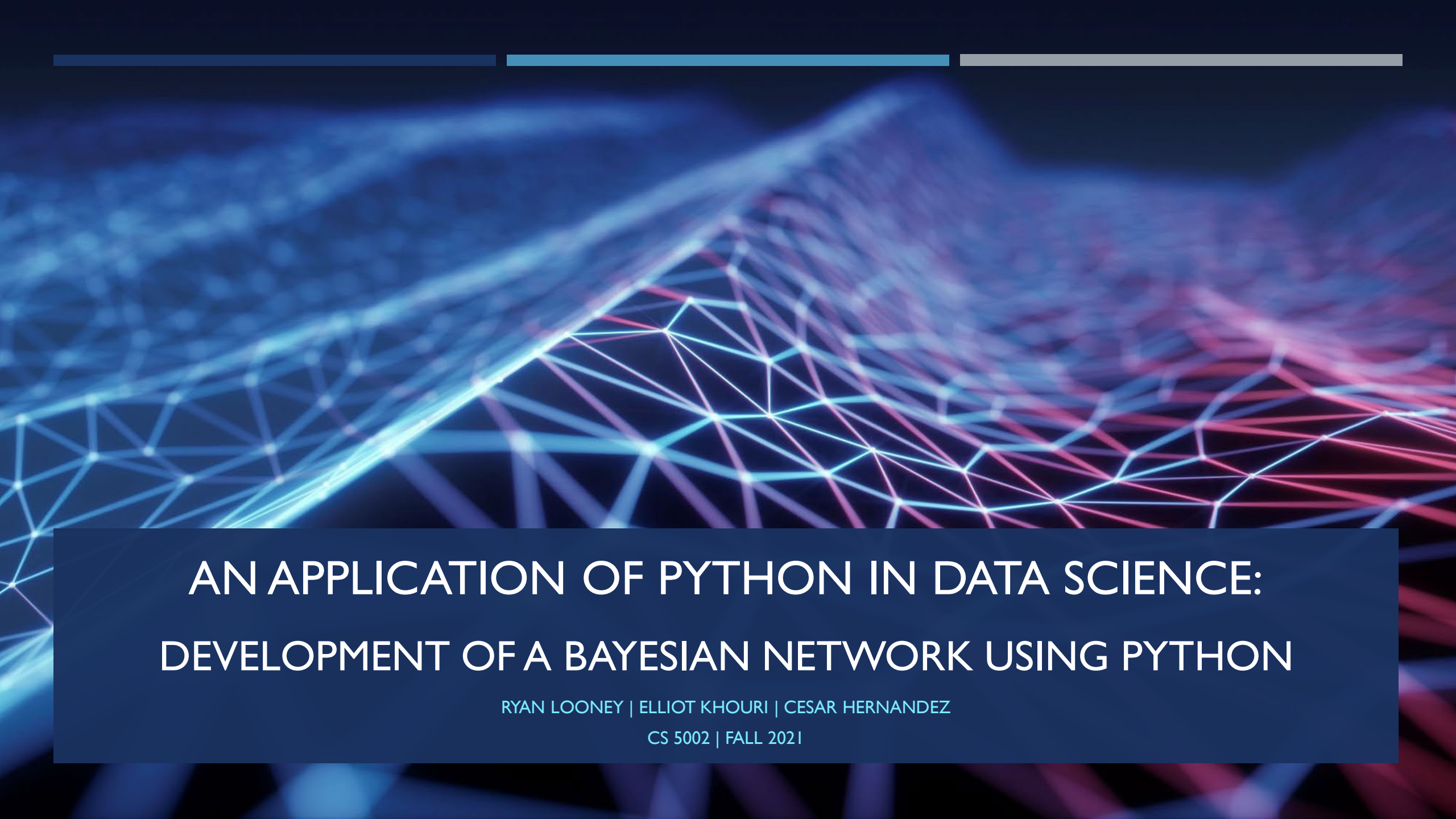


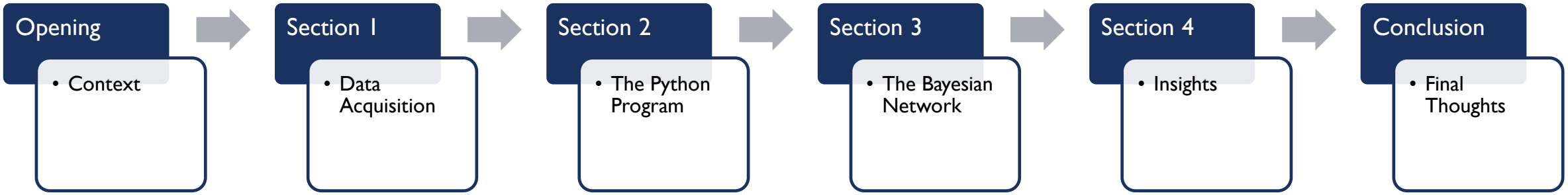
AN APPLICATION OF PYTHON IN DATA SCIENCE:

DEVELOPMENT OF A BAYESIAN NETWORK USING PYTHON

RYAN LOONEY | ELLIOT KHOURI | CESAR HERNANDEZ

CS 5002 | FALL 2021

SCHEDULE


Opening
• Context
Section 1
• Data Acquisition
Section 2
• The Python Program
Section 3
• The Bayesian Network
Section 4
• Insights
Conclusion
• Final Thoughts

OPENING

CONTEXT & RELEVANCE

CONTEXT (SCOPE)

Our Focus

1. Scoring First Advantage
2. Quarter Lead Advantage
3. Previous Record Advantage

Out of Our focus

Countless factors that can affect winning…

1. Injuries
2. Rival Matchups
3. Drafting
4. Team Chemistry
5. Team Experience
6. Fatigue
7. Overtime
8. Playoffs

CONTEXT (QUESTION & HYPOTHESIS)

Question

1. Does scoring first give an advantage?
2. Does leading in the 1st, 2nd , 3rd or 4th quarter give an advantage?
3. Does previous record give an advantage?

(Note: By advantage we mean a greater chance of winning.)

Hypothesis

1. If a team scores first then we believe it gives an advantage.
2. If a team leads in a quarter then we believe it gives an advantage (compounding).
3. If a team has a better previous record, then we believe it gives an advantage

RELEVANCE

Why is this important to us…

SECTION 1.

DATA ACQUISITION

DATA ACQUISITION (FINDING THE RAW DATA)

kaggle

DATA ACQUISITION (UNDERSTANDING THE RAW DATA)

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
URL	GameTyp	Location	Date	Time	WinningT	Quarter	SecLeft	AwayTear	AwayPlay	AwayScor	HomeTea	HomePlay	HomeScor	Shooter	ShotType	ShotOutc	ShotDist	Assister	Blocker	FoulType	Fouler	Fouled	Rebounde	ReboundT	ViolationF	ViolationT	TimeoutT	FreeThrov
/boxscore	regular	Philips Ar	October 2	8:00 PM	DET	1	720	DET	Jump ball:	0	ATL		0															
/boxscore	regular	Philips Ar	October 2	8:00 PM	DET	1	701	DET	A. Drumm	0	ATL		0	A. Drumm	2-pt layup	miss	1		A. Horford	- horfoal01								
/boxscore	regular	Philips Ar	October 2	8:00 PM	DET	1	699	DET		0	ATL	Defensive	0										K. Bazemc	defensive					
/boxscore	regular	Philips Ar	October 2	8:00 PM	DET	1	697	DET		0	ATL	Turnover	0																
/boxscore	regular	Philips Ar	October 2	8:00 PM	DET	1	681	DET	M. Morris	2	ATL		0	M. Morris	2-pt jump	make	13	A. Drummond	- drumman01										
/boxscore	regular	Philips Ar	October 2	8:00 PM	DET	1	660	DET		2	ATL	P. Millsap	2	P. Millsap	2-pt jump	make	12	K. Korver	- korveky01										
/boxscore	regular	Philips Ar	October 2	8:00 PM	DET	1	644	DET	K. Caldwe	4	ATL		2	K. Caldwe	2-pt jump	make	8	E. Ä°lyasova	- ilyaser01										
/boxscore	regular	Philips Ar	October 2	8:00 PM	DET	1	627	DET		4	ATL	A. Horforc	2	A. Horforc	2-pt jump	miss	20												
/boxscore	regular	Philips Ar	October 2	8:00 PM	DET	1	625	DET	Defensive	4	ATL		2										R. Jackson	defensive					

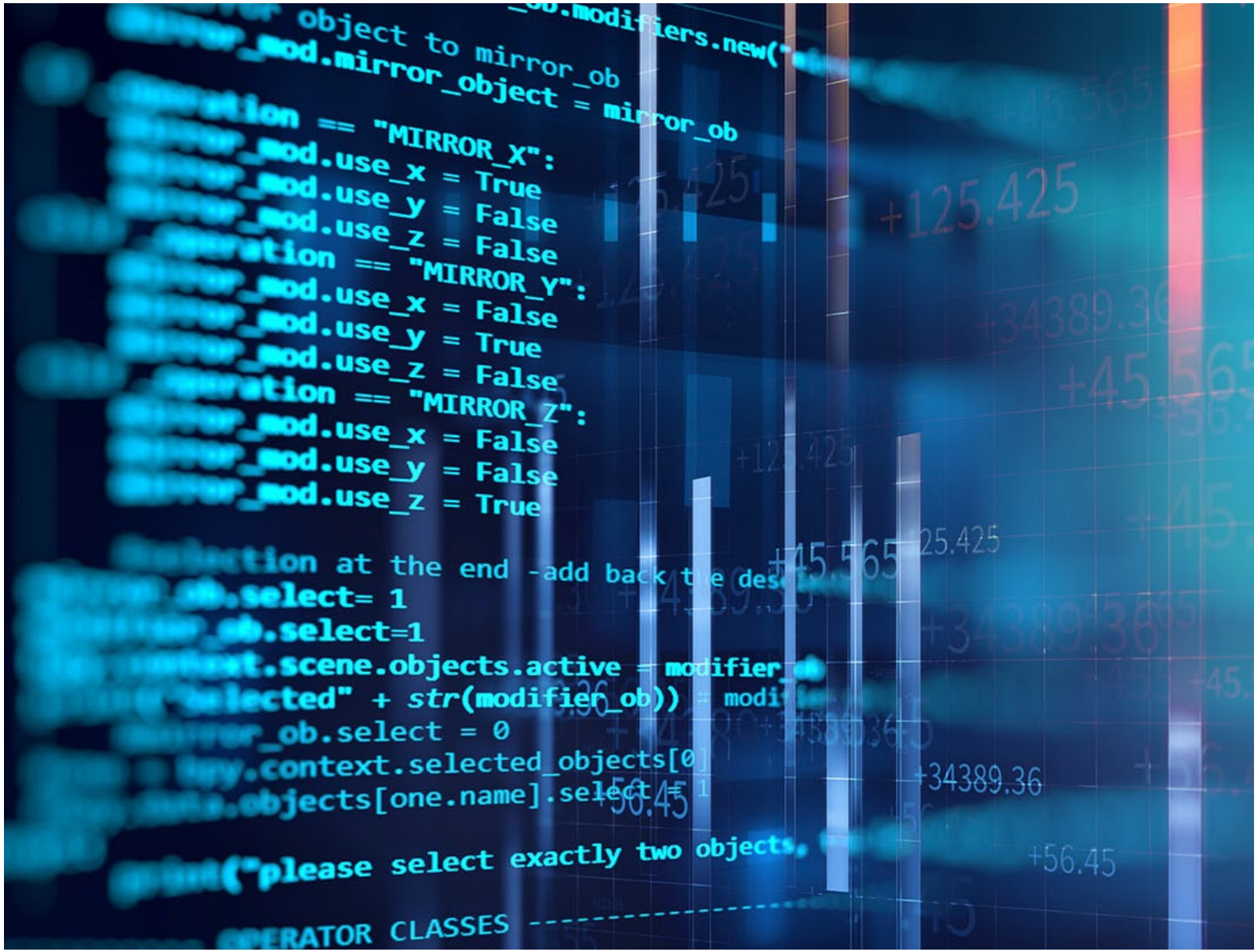
L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN
HomeTea	HomePlay	HomeScor	Shooter	ShotType	ShotOutc	ShotDist	Assister	Blocker	FoulType	Fouler	Fouled	Rebounde	ReboundT	ViolationF	ViolationT	TimeoutT	FreeThrov	FreeThrov	FreeThrov	EnterGam	LeaveGam	TurnoverF	TurnoverT	TurnoverC	TurnoverC	JumpballA	JumpballH	JumpballPos
ATL		0																								A. Drumm	A. Horforc	E. Ä°lyasova
ATL		0	A. Drumm	2-pt layup	miss	1		A. Horford	- horfoal01																			
ATL	Defensive	0										K. Bazemc	defensive															
ATL	Turnover	0																				K. Bazemc	bad pass					
ATL		0	M. Morris	2-pt jump	make	13	A. Drummond	- drumman01																				
ATL	P. Millsap	2	P. Millsap	2-pt jump	make	12	K. Korver	- korveky01																				
ATL		2	K. Caldwe	2-pt jump	make	8	E. Ä°lyasova	- ilyaser01																				
ATL	A. Horforc	2	A. Horforc	2-pt jump	miss	20																						

DATA ACQUISITION (THE NEED FOR PYTHON)

1. Needed to convert raw data from play/play to game format
2. Needed to determine the home team quarterly status
3. Needed to determine the away team quarterly status
4. Needed to determine which team scored first

SECTION 2.

THE PYTHON PROGRAM



THE PYTHON PROGRAM (EXPLANATION OF THE CODE)

```
            elif is_end_of_second and home_score < away_score:
                games_dict[game_id].update({'home_leads_after_second' : False})
                games_dict[game_id].update({'away_leads_after_second' : True})
            elif is_end_of_second and home_score == away_score:
                games_dict[game_id].update({'home_leads_after_second' : 'Tie'})
                games_dict[game_id].update({'away_leads_after_second' : 'Tie'})


            ###Does the home team lead after third?
            is_end_of_third = (quarter == 3 and home_play == 'End of 3rd quarter' or away_play == 'End of 3rd quarter')

            if is_end_of_third and home_score > away_score:
                games_dict[game_id].update({'home_leads_after_third' : True})
                games_dict[game_id].update({'away_leads_after_third' : False})
            elif is_end_of_third and home_score < away_score:
                games_dict[game_id].update({'home_leads_after_third' : False})
                games_dict[game_id].update({'away_leads_after_third' : False})
            elif is_end_of_third and home_score == away_score:
                games_dict[game_id].update({'home_leads_after_third' : 'Tie'})
                games_dict[game_id].update({'away_leads_after_third' : 'Tie'})


            ###Does the home team lead during the fourth?
            is_middle_of_fourth = (quarter == 4 and (seconds_left > 240 and seconds_left <= 480))
            if is_middle_of_fourth and home_score > away_score:
                games_dict[game_id].update({'home_leads_mid_fourth' : True})
                games_dict[game_id].update({'away_leads_mid_fourth' : False})
            elif is_middle_of_fourth and home_score < away_score:
                games_dict[game_id].update({'home_leads_mid_fourth' : False})
                games_dict[game_id].update({'away_leads_mid_fourth' : True})
            elif is_middle_of_fourth and home_score == away_score:
                games_dict[game_id].update({'home_leads_mid_fourth' : 'Tie'})
                games_dict[game_id].update({'away_leads_mid_fourth' : 'Tie'})

    #Create a new data frame from the dictionary created in the for loop
    games_frame = pd.DataFrame.from_dict(games_dict, orient='index')
    games_frame.to_excel("games_out.xlsx")
    return games_frame
```

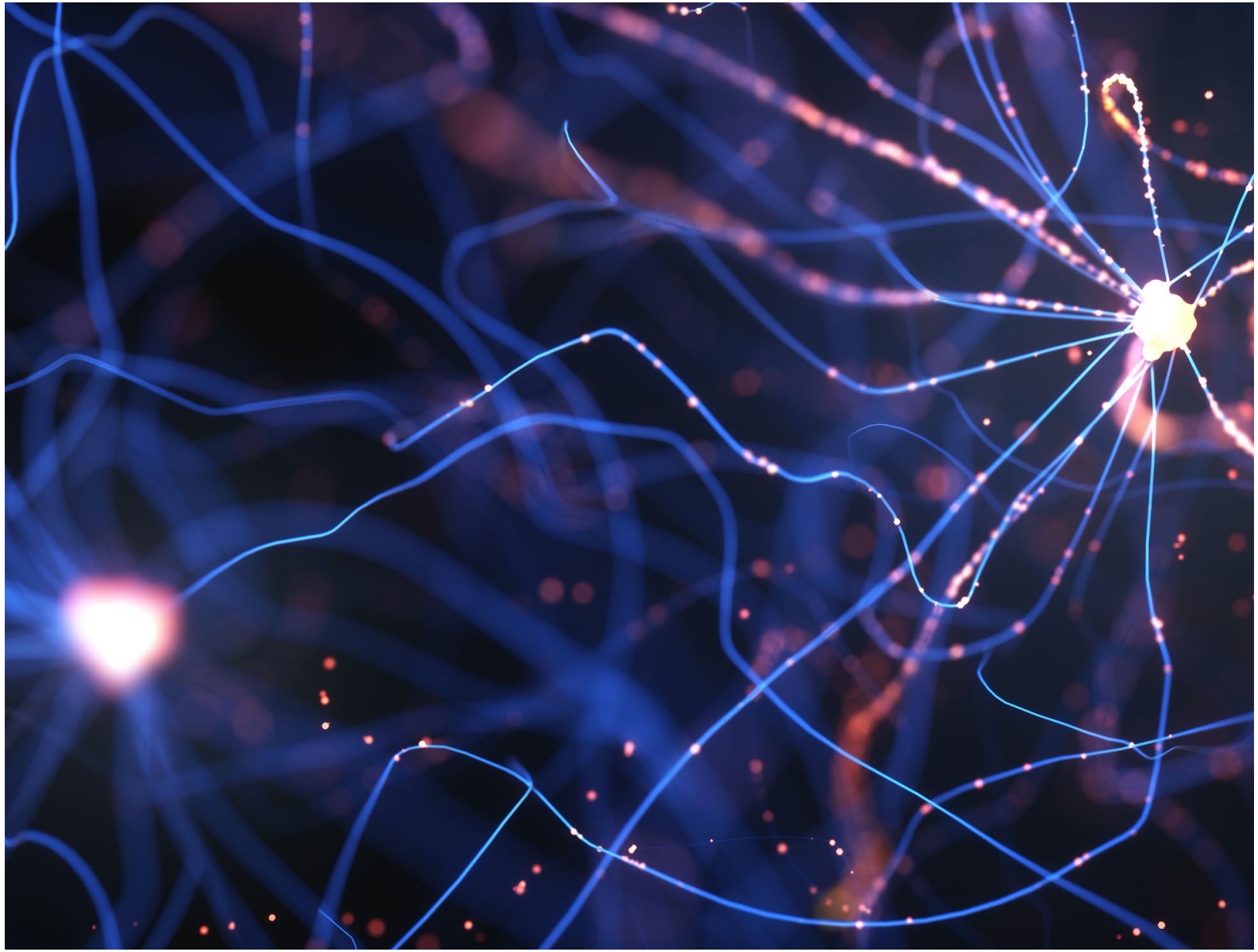
THE PYTHON PROGRAM (THE OUTPUT)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
		date	home_team	away_team	winning_team	home_team_wins	away_team_wins	home_team_scores_first	away_team_scores_first	home_leads_after_first	away_leads_after_first	home_leads_after_second	away_leads_after_second	home_leads_after_third
1	201510270ATL	October 27 2015	ATL	DET	DET	FALSE	TRUE	FALSE	TRUE	Tie	Tie	FALSE	TRUE	FALSE
2	201510270CHI	October 27 2015	CHI	CLE	CHI	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE
3	201510270GSW	October 27 2015	GSW	NOP	GSW	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE
4	201510280ORL	October 28 2015	ORL	WAS	WAS	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE
5	201510280BOS	October 28 2015	BOS	PHI	BOS	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
6	201510280BRK	October 28 2015	BRK	CHI	CHI	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE
7	201510280DET	October 28 2015	DET	UTA	DET	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE
8	201510280MIA	October 28 2015	MIA	CHO	MIA	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
9	201510280TOR	October 28 2015	TOR	IND	TOR	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE
10	201510280HOU	October 28 2015	HOU	DEN	DEN	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE
11	201510280MEM	October 28 2015	MEM	CLE	CLE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE
12	201510280MIL	October 28 2015	MIL	NYK	NYK	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE

N	O	P	Q
home_leads_after_third	away_leads_after_third	home_leads_mid_fourth	away_leads_mid_fourth
FALSE	TRUE	FALSE	TRUE
TRUE	FALSE	TRUE	FALSE
TRUE	FALSE	TRUE	FALSE
TRUE	FALSE	FALSE	TRUE
TRUE	FALSE	TRUE	FALSE
FALSE	TRUE	FALSE	TRUE
TRUE	FALSE	TRUE	FALSE
TRUE	FALSE	TRUE	FALSE
TRUE	FALSE	TRUE	FALSE
FALSE	TRUE	FALSE	TRUE
FALSE	TRUE	FALSE	TRUE

SECTION 3.

THE BAYESIAN NETWORK

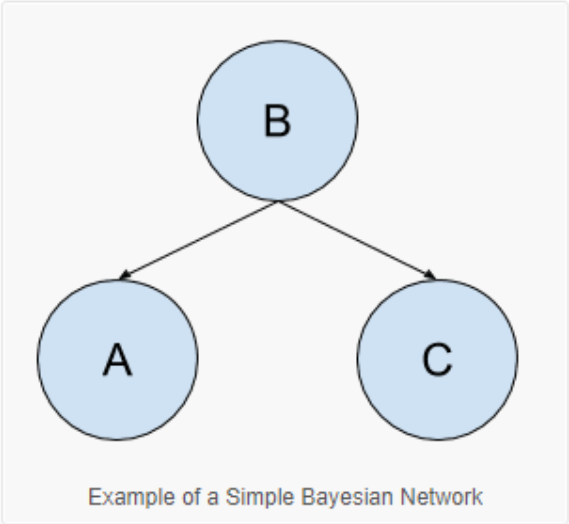


BAYESIAN NETWORK OVERVIEW

The model summarizes the joint probability of P(A, B, C), calculated as:

- P(A, B, C) = P(A|B) * P(C|B) * P(B)

We can draw the graph as follows:



Example of a Simple Bayesian Network

- Graphical model with directed edges and no cycles.
- "Bayesian network models capture both conditionally dependent and conditionally independent relationships between random variables." (Brownlee)
- "Models can be prepared by experts or learned from data, then used for inference to estimate the probabilities for causal or subsequent events." (Brownlee)

THE BAYESIAN NETWORK (THE PREDICTIVE BAYESIAN MODEL)

1. The Predictive Bayesian Model
2. The Exhaustive Bayesian Model
3. The Conditional Probability Model

SECTION 4.

INSIGHTS

INSIGHTS

There is a clear home team advantage that implies a greater chance of winning.

1. If the home team is winning in the 4th quarter, they will win 92% of the time versus the away team's 88% under the same circumstances
2. This discrepancy in probability percentage evident for all quarters, and widens the farther away the away team holds a lead from the 4th quarter
    a) The away team wins ~59% of the time that they are leading the first quarter compared to ~72% for the Home Team
3. Even if the away team wins every quarter up to the middle of the 4th quarter, they only have a 92% chance of winning compared to a 97% chance for the home team in the same circumstances.

INSIGHTS(CONTINUED)

Leading in each subsequent quarter gives a compounding increase to chance of victory.

a) Chance of lead diminishes for either team when the lead flip-flops quarter to quarter

b) Best chance for either team to win comes when they lead all quarters

c) Under same circumstances the trend is evident for away team with lessened effect

Pre-conditions to outcome	Home Team wins
ATSF & HL1	70.59%

Pre-conditions to outcome	Home Team wins
HTSF & HL1 & HL2	84.09%

Pre-conditions to outcome	Home Team wins
HTSF & HL1 & HL2 & HL3	92.64%

Pre-conditions to outcome	Home Team Wins
HTSF & HL1 & HL2 & HL3 & HL4	97.70%

INSIGHTS(CONTINUED)

Scoring first implies a greater chance of winning, but it is less impactful than leading quarters

Pre-conditions to outcome	Home Team wins
HTSF & HL1 & HL2	84.09%
ATSF & HL1 & HL2	78.26%

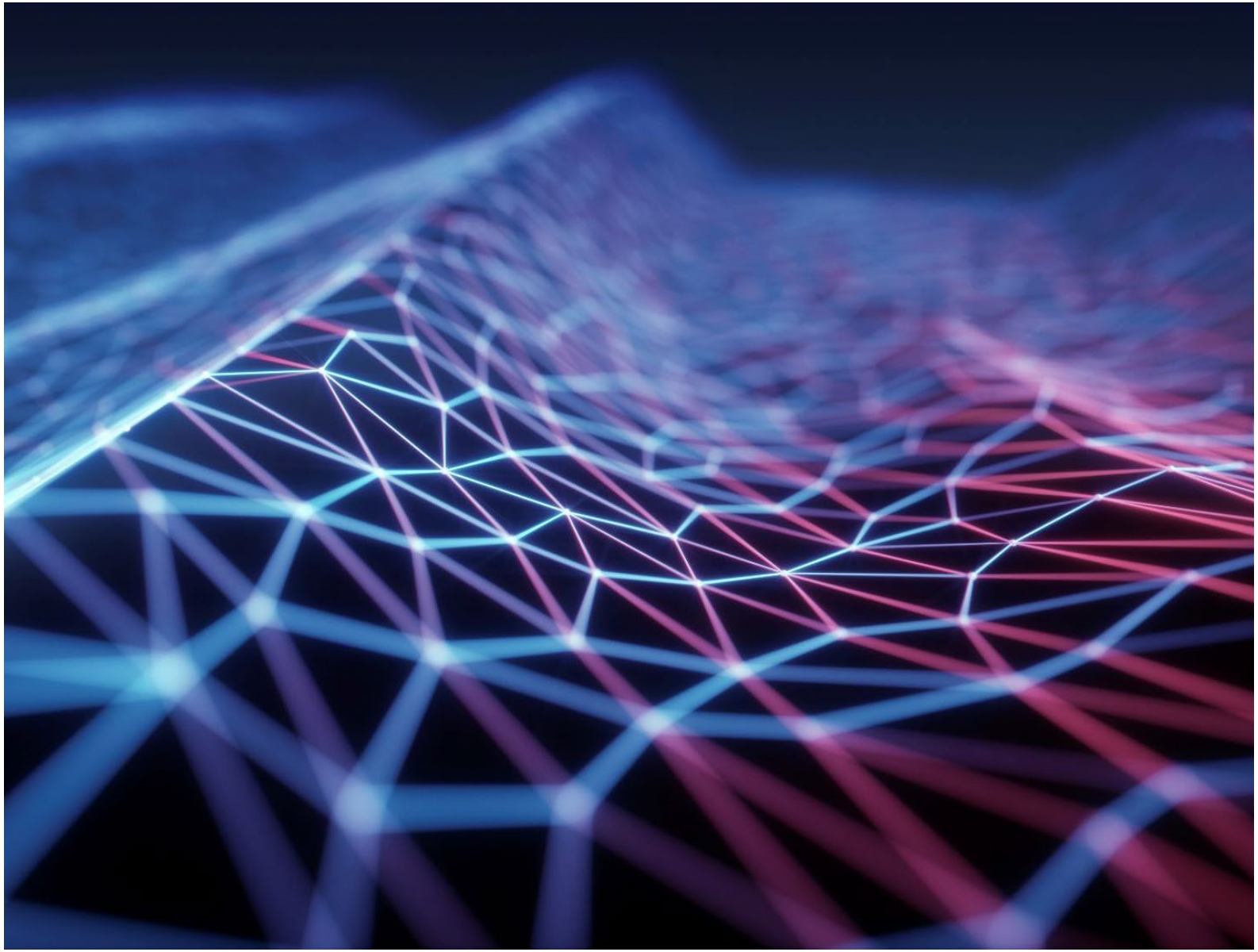
All else held equal in this case, the away team scoring 1st drops the home team's victory chance by 7%

Pre-conditions to outcome	Away Team Wins
ATSF & AL1 & AL2	73.60%
HTSF & AL1 & AL2	71.68%

All else held equal in this case, the home team scoring 1st drops the away team's victory chance by ~2%

SECTION 5.

CONCLUSION



REVISITING THE QUESTION AND HYPOTHESES

Does scoring first imply a greater chance of winning?

Our data shows yes, though it is marginal

Does leading in the 1st, 2nd , 3rd or 4th imply a greater chance of winning?

Data bears out yes, and that this advantage is significant the more consistently that a lead is held.

Does previous record imply a greater chance of winning?

Remains unexplored due to time limitations for now – would be a good step to add this variable to the network.

WEAKNESSES AND LIMITATIONS

Primary Factors

1. Previous Record Advantage
2. Playoff Games & Overtime Games
3. Back to Back Games (Fatigue)

Secondary Factors

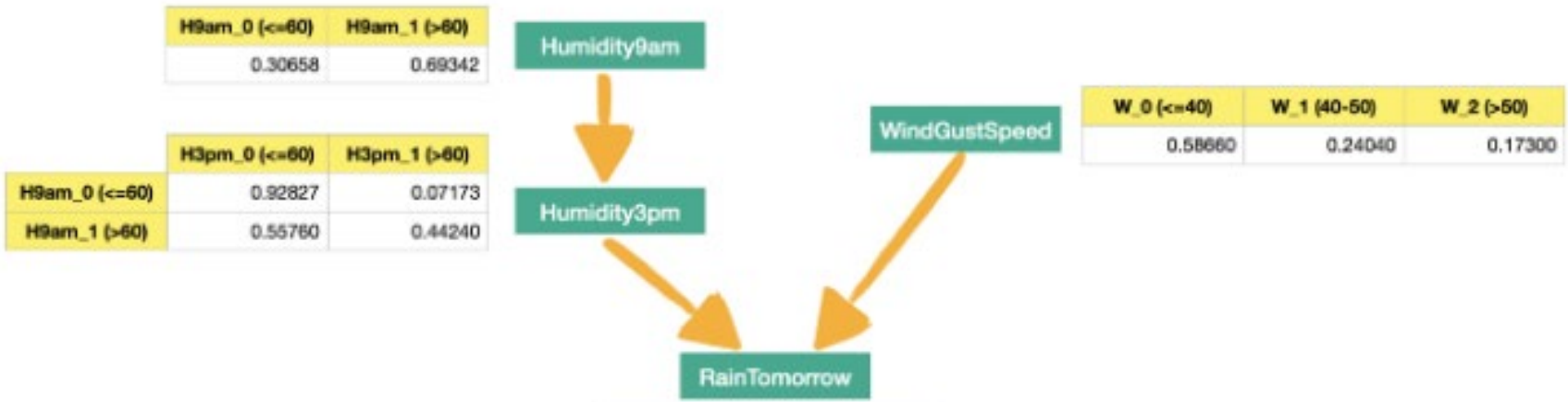
1. Injuries
2. Rival Matchups
3. Drafting
4. Team Chemistry
5. Team Experience

FUTURE INITIATIVES

Creating a more robust "network" vs. a chain…

1. Previous Record Advantage
2. Playoff Games & Overtime Games
3. Back to Back Games (Fatigue)

FUTURE INITIATIVES(CONTINUED)



H9am_0 (<=60)	H9am_1 (>60)
0.30658	0.69342

	H3pm_0 (<=60)	H3pm_1 (>60)
H9am_0 (<=60)	0.92827	0.07173
H9am_1 (>60)	0.55760	0.44240

W_0 (<=40)	W_1 (40-50)	W_2 (>50)
0.58660	0.24040	0.17300

	RT_0 (No)	RT_1 (Yes)
H3pm_0 (<=60), W_0 (<=40)	0.92314	0.07686
H3pm_0 (<=60), W_1 (40-50)	0.89072	0.10928
H3pm_0 (<=60), W_2 (>50)	0.76008	0.23992
H3pm_1 (>60), W_0 (<=40)	0.64250	0.35750
H3pm_1 (>60), W_1 (40-50)	0.49168	0.50832
H3pm_1 (>60), W_2 (>50)	0.32182	0.67818


End

SOURCES REFERENCED THROUGHOUT PROJECT

Bayesian Network Theory

https://www.edureka.co/blog/bayesian-networks/

https://towardsdatascience.com/bbn-bayesian-belief-networks-how-to-build-them-effectively-in-python-6b7f93435bba

https://cs.calvin.edu/courses/cs/344/kvlinden/resources/AIMA-3rd-edition.pdf

https://machinelearningmastery.com/introduction-to-bayesian-belief-networks/

https://machinelearningmastery.com/introduction-to-bayesian-belief-networks/ (Information, quotes, and graphic from Jason Brownlee's work on this page.)

Pandas Dataframe Documentation (for play-by-play data transformation)

https://www.geeksforgeeks.org/python-pandas-dataframe/