

United States



13

FT

0




Thailand



U.S. makes history in its World Cup opener

Alex Morgan scored a record-tying five goals as the USWNT notched the biggest blowout in World Cup history, demolishing Group F opponent Thailand.

Soccer Analytics



“Data is like garbage. You’d
better know what you are going
to do with it before you collect it.”
Mark Twain

What kind of garbage do we want ?

Hypotheses

1. **Being a Home Team gives better % chance of winning**
2. **Being home increases a team's odds of winning according to bookmakers**
3. **Higher defensive aggression leads to higher winning percentage**
4. **Higher offensive attributes leads to higher winning percentage**

Data Used;

Kaggle European Soccer Data set

- **2008-2016 Years**
- **483 Teams**
- **25,979 Matches**

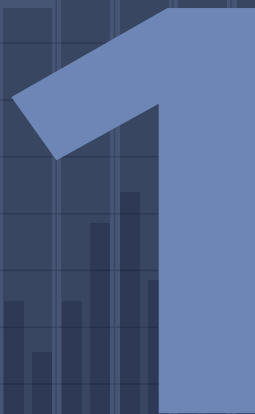
Approach

1. Query what we need
2. Examine Data
3. Visualize Data
4. Test our Hypotheses
5. Find significance or non significance
6. Reject or Accept our Null



Searching and Cleaning Garbage

- Finding all the Home wins and Away wins for each team, with the amount of home and away games played.
 - Challenges were that the data set only had scores of each game matched with teams api numbers
- Realized something's not right since this one Polish Team keeps messing with your data since they had multiple fallouts due to corruption and bankruptcy
 - Finally was able to remove them entirely after finding out they also had two different team api numbers with the same name.



Slide into those DFs

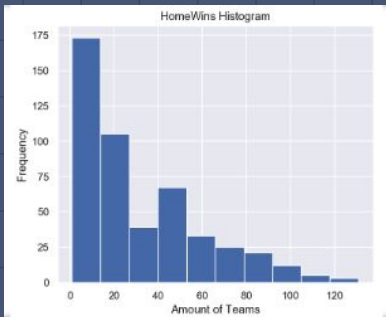
- Dataframes:
 - Home(Wins,Draws,#ofgames,Team)
 - (%Wins...Difference of Wins)
 - Away(Wins,Draws,#ofgames,Team)
 - (%Wins...Difference of Wins)
 - Merged Them

```
In [288]: CombinedHomeandAway[CombinedHomeandAway['team_long_name']=='Arsenal']
```

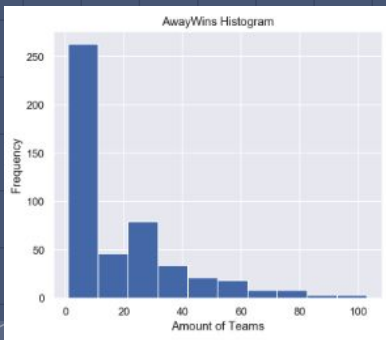
```
Out[288]:
```

ing_name	HomeDraws	Home_Wins	NumOfHomeGames	AwayDraws	Team.team_api_id	Away_Wins	NumOfAwayGames	%HomeWinSuccess	%
Arsenal	34	97	152	39	9825	73	152	0.638158	

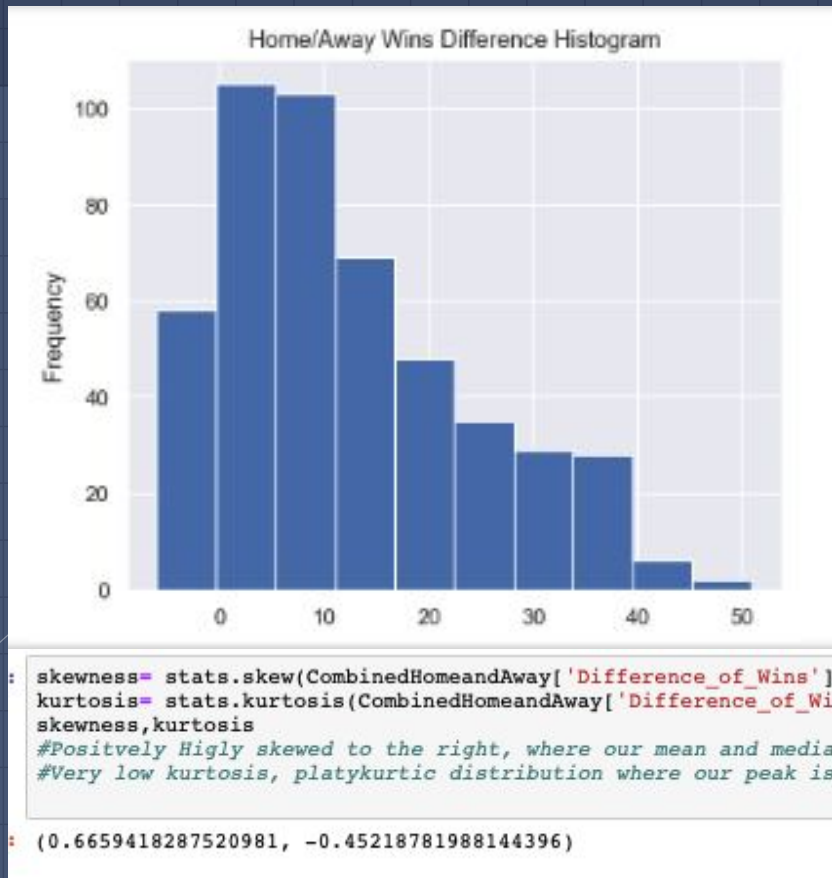
Clues for testing



Left two: their okay but don't say much



Right: Clearly see that not even 60 teams won more away games than home. Allows us to test further.



1st Hypothesis:

ALT: There's a statistical difference between Winning Home games than Away games

Null: They're statistically the same.

T-Test:

Checking whether they differ significantly across their games. Our very small P-value and large T-stat shows that *our Null hypothesis is False and can be rejected.*



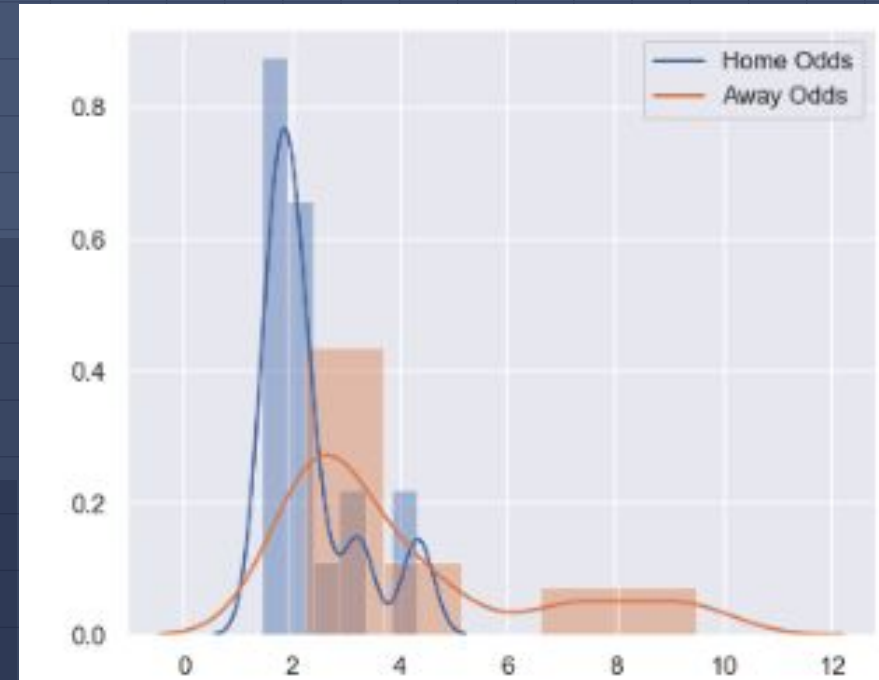
```
In [124]: ttest_rel(CombinedHomeandAway['Home_Wins'], CombinedHomeandAway['Away_Wins'])  
#Repeated T-Test to account for using the same team twice for home and away. Measure  
#Checking whether they differ significantly across their games. Our very small p-value  
#Null hypothesis (Home and away wins have no statistical difference in averages)
```

```
Out[124]: Ttest_relResult(statistic=22.948667608665446, pvalue=2.5957348349317354e-79)
```


2nd Hypothesis: Is there a Statistical Difference for a Home Team in terms of the Odds?

- **The bookmakers' numbers say yes!**
- The lower the odds, the higher chance the team has of winning
 - Graph on right displays Manchester City's odds against all 19 opponents, both home and away
 - When Manchester City was home, it often had lower odds than when they were away

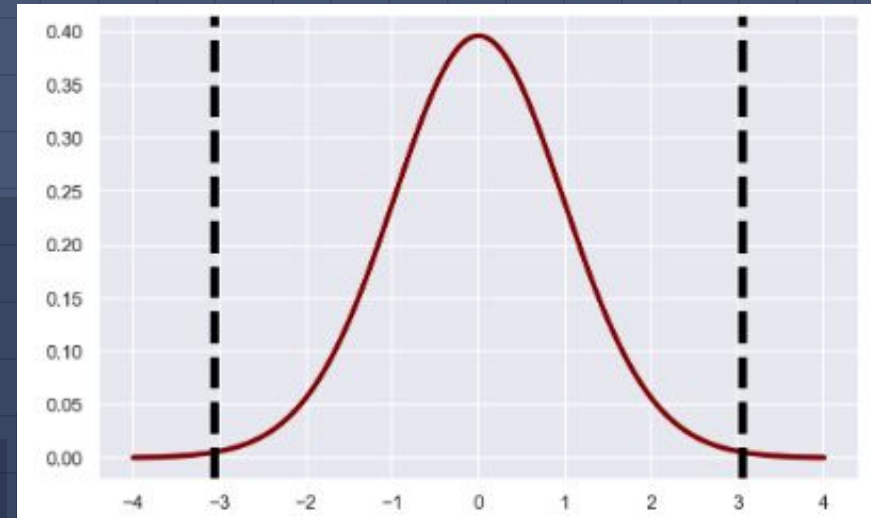
Manchester City 2008/2009 Bet 365 Odds



2nd Hypothesis: Is there a Statistical Difference for a Home Team in terms of the Odds?

- If any of the odds fell outside of the region marked in the black dotted line, we would have to assume that there is no difference between Home Odds and Away Odds
 - However, this is not the case, with a p-value of 0.00416, we can reject our null hypothesis and accept our alternative as a fact.

T-Distribution for our Home and Away Odds



3rd Hypothesis: Does a high defensive aggression rating from FIFA video game mean a higher winning % in real life?

- **The numbers (shockingly) say no!**
- We (read: Kaggle) downloaded FIFA game data for each season starting in 2010
 - For team defense, the attributes measured were:
 - Defense Pressure
 - Defense Aggression
 - Defense Team Width
- We then calculated the winning % of each team each season and tried to find a correlation between the two

	Wins	Loss	Draws	Winning %	Defense Pressure	Defense Aggression	Defense Team Width	Build Up Speed	Build Up Passing	Chance Creation Passing	Chance Creation Crossing	Chance Creation Shooting
Wins	1.00	-6.85e-01	-0.30	9.22e-01	0.15	0.09	0.19	0.24	5.13e-02	0.25	1.87e-01	0.30
Loss	-0.68	1.00e+00	-0.07	-9.05e-01	-0.05	0.02	-0.04	-0.01	7.57e-02	-0.12	-8.00e-03	-0.04
Draws	-0.30	-6.80e-02	1.00	-1.32e-01	-0.08	0.03	-0.05	0.04	9.35e-02	-0.04	2.26e-02	-0.22
Winning %	0.92	-9.05e-01	-0.13	1.00e+00	0.11	0.05	0.13	0.14	-9.76e-04	0.22	1.17e-01	0.20
Defense Pressure	0.15	-4.74e-02	-0.08	1.14e-01	1.00	0.69	0.88	0.55	4.83e-01	0.60	3.78e-01	0.57
Defense Aggression	0.09	2.45e-02	0.03	4.99e-02	0.69	1.00	0.67	0.57	6.58e-01	0.60	4.84e-01	0.46
Defense Team Width	0.19	-3.57e-02	-0.05	1.28e-01	0.88	0.67	1.00	0.55	4.73e-01	0.58	4.01e-01	0.59
Build Up Speed	0.24	-1.22e-02	0.04	1.41e-01	0.55	0.57	0.55	1.00	6.59e-01	0.66	6.13e-01	0.47
Build Up Passing	0.05	7.57e-02	0.09	-9.76e-04	0.48	0.66	0.47	0.66	1.00e+00	0.57	6.01e-01	0.37
Chance Creation Passing	0.25	-1.18e-01	-0.04	2.16e-01	0.60	0.60	0.58	0.66	5.66e-01	1.00	6.62e-01	0.59
Chance Creation Crossing	0.19	-8.00e-03	0.02	1.17e-01	0.38	0.48	0.40	0.61	6.01e-01	0.66	1.00e+00	0.50
Chance Creation Shooting	0.30	-4.10e-02	-0.22	2.00e-01	0.57	0.46	0.59	0.47	3.69e-01	0.59	5.01e-01	1.00



4th Hypothesis: If (fake video game) defense doesn't win championships, does (fake video game) offense?

- **The numbers (shockingly) say (kind of) YES!**
- Using the same data we (read: Kaggle) downloaded from FIFA, the team offensive metrics we obtained were:
 - Chance Creation Passing
 - Chance Creation Crossing
 - Chance Creation Shooting
- Using the data that we calculated earlier, we ran the same correlation table

	Wins	Loss	Draws	Winning %	Defense Pressure	Defense Aggression	Defense Team Width	Build Up Speed	Build Up Passing	Chance Creation Passing	Chance Creation Crossing	Chance Creation Shooting
Wins	1.00	-6.85e-01	-0.30	9.22e-01	0.15	0.09	0.19	0.24	5.13e-02	0.25	1.87e-01	0.30
Loss	-0.68	1.00e+00	-0.07	-9.05e-01	-0.05	0.02	-0.04	-0.01	7.57e-02	-0.12	-8.00e-03	-0.04
Draws	-0.30	-6.80e-02	1.00	-1.32e-01	-0.08	0.03	-0.05	0.04	9.35e-02	-0.04	2.26e-02	-0.22
Winning %	0.92	-9.05e-01	-0.13	1.00e+00	0.11	0.05	0.13	0.14	-9.76e-04	0.22	1.17e-01	0.20
Defense Pressure	0.15	-4.74e-02	-0.08	1.14e-01	1.00	0.69	0.88	0.55	4.83e-01	0.60	3.78e-01	0.57
Defense Aggression	0.09	2.45e-02	0.03	4.99e-02	0.69	1.00	0.67	0.57	6.58e-01	0.60	4.84e-01	0.46
Defense Team Width	0.19	-3.57e-02	-0.05	1.28e-01	0.88	0.67	1.00	0.55	4.73e-01	0.58	4.01e-01	0.59
Build Up Speed	0.24	-1.22e-02	0.04	1.41e-01	0.55	0.57	0.55	1.00	6.59e-01	0.66	6.13e-01	0.47
Build Up Passing	0.05	7.57e-02	0.09	-9.76e-04	0.48	0.66	0.47	0.66	1.00e+00	0.57	6.01e-01	0.37
Chance Creation Passing	0.25	-1.18e-01	-0.04	2.16e-01	0.60	0.60	0.58	0.66	5.66e-01	1.00	6.62e-01	0.59
Chance Creation Crossing	0.19	-8.00e-03	0.02	1.17e-01	0.38	0.48	0.40	0.61	6.01e-01	0.66	1.00e+00	0.50
Chance Creation Shooting	0.30	-4.10e-02	-0.22	2.00e-01	0.57	0.46	0.59	0.47	3.69e-01	0.59	5.01e-01	1.00

4th Hypothesis: If (fake video game) defense doesn't win championships, does (fake video game) offense?

- **While the correlations were relatively low for all 3 categories, the p-values for both “Chance Creation Passing” and “Chance Creation Shooting” resulted in p-values less than .05, meaning we had to reject our null hypothesis.**
 - “Chance Creation Crossing” did not have a p-value below .05, therefore we did not reject the null hypothesis

```
In [77]: scipy.stats.pearsonr(Season_2010['Winning %'], Season_2010['Chance Creation Passing'])
```

```
Out[77]: (0.2157968300728158, 0.0031767136077806854)
```

```
In [78]: scipy.stats.pearsonr(Season_2010['Winning %'], Season_2010['Chance Creation Crossing'])
```

```
Out[78]: (0.11657251970332994, 0.11406032184266258)
```

```
In [79]: scipy.stats.pearsonr(Season_2010['Winning %'], Season_2010['Chance Creation Shooting'])
```

```
Out[79]: (0.20033994291514065, 0.006252732733704142)
```


Key Statistical Takeaways

- There is a higher percent chance of winning when you are the home team than the away team
- The bookmakers believe that Home Teams have a better chance of winning than away, and therefore give lower odds to the Home Team if everything else remains the same
- Fake defensive statistics from video games **DO NOT** translate to a higher real life winning %
- Fake offensive statistics from videos games **DO (kind of)** translate to a higher real life winning %

THANKS!

Any questions?



For the Boys in Blue!