Data Collection

Before building the simulator, we realized that a lot of data was needed to ensure its accuracy. Thus, we scoured the internet for relevant datasets and useful information. The following section presents a summary of the data found, including sources.

European Soccer Database

This dataset on Kaggle.com came in the form of an SQL database with 7 tables. These contained data from 10 major soccer leagues around the world, including the English Premier League.

Table	Total Rows	Total Columns	Columns
Country	11	2	id, name
League	11	3	id, country_id, name
Match	25979	115	id, country_id, league_id, season, stage, date, match_api_id, home_team_api_id, away_team_api_id, home_team_goal, away_team_goal, home_player_X1, home_player_X2, home_player_X3, home_player_X4, home_player_X5, home_player_X6, home_player_X7, home_player_X8, home_player_X9, home_player_X10, home_player_X11, away_player_X1, away_player_X2, away_player_X3, away_player_X4, away_player_X5, away_player_X6, away_player_X7, away_player_X8, away_player_X9, away_player_X10, away_player_X11, home_player_Y1, home_player_Y2, home_player_Y3, home_player_Y4, home_player_Y5, home_player_Y6, home_player_Y7, home_player_Y8, home_player_Y9, home_player_Y10, home_player_Y11, away_player_Y1, away_player_Y2, away_player_Y3, away_player_Y4, away_player_Y5, away_player_Y6, away_player_Y7, away_player_Y8, away_player_Y9, away_player_Y10, away_player_Y11, home_player_1, home_player_2, home_player_3, home_player_4, home_player_5, home_player_6, home_player_7, home_player_8, home_player_9, home_player_10, home_player_11, away_player_1, away_player_2, away_player_3, away_player_4, away_player_5, away_player_6, away_player_7, away_player_8, away_player_9, away_player_10, away_player_11, goal, shoton, shotoff, foulcommit, card, cross, corner, possession, B365H, B365D, B365A, BWH, BWD, BWA, IWH, IWD, IWA, LBH, LBD, LBA, PSH, PSD, PSA, WHH, WHD, WHA, SJH, SJD, SJA, VCH, VCD, VCA, GBH, GBD, GBA, BSH, BSD, BSA

The above tables provided most of the real life data we required, in particular the Match table. This table has statistics for number of goals scored by the home team and away team in each match from 2008 to 2016. The last few columns also included the bookie predictions and odds for each match. Unfortunately, the in game statistics such as number of shots, possession and others were corrupted, showing nonsensical strings such as this:

goal	shoton	shotoff	foulcommit	card	cross	corner	possession
<goal><value><commer	<shoton><value><stats	<shotoff><value><stats	<foulcommit><value><s	<card><value><commer	<cross><value><stats><	<corner><value><stats>	<possession><value><c
<goal><value><commer	<shoton><value><stats	<shotoff><value><stats	<foulcommit><value><s	<card><value><commer	<cross><value><stats><	<corner><value><stats>	<possession><value><c
<goal><value><commer	<shoton><value><stats	<shotoff><value><stats	<foulcommit><value><s	<card><value><commer	<cross><value><stats><	<corner><value><stats>	<possession><value><c
<goal><value><commer	<shoton><value><event	<shotoff><value><stats	<foulcommit><value><s	<card />	<cross><value><stats><	<corner><value><stats>	<possession><value><c
<goal><value><commer	<shoton><value><stats	<shotoff><value><stats	<foulcommit><value><s	<card />	<cross><value><event_i	<corner><value><stats>	<possession><value><c
<goal><value><commer	<shoton><value><stats	<shotoff><value><stats	<foulcommit><value><s	<card><value><commer	<cross><value><stats><	<corner><value><stats>	<possession><value><c
<goal><value><commer	<shoton><value><stats	<shotoff><value><stats	<foulcommit><value><s	<card><value><commer	<cross><value><stats><	<corner><value><stats>	<possession><value><c
<goal><value><commer	<shoton><value><stats	<shotoff><value><stats	<foulcommit><value><s	<card><value><commer	<cross><value><stats><	<corner><value><stats>	<possession><value><c
<goal><value><commer	<shoton><value><stats	<shotoff><value><stats	<foulcommit><value><s	<card><value><commer	<cross><value><event_i	<corner><value><stats>	<possession><value><c
<goal><value><commer	<shoton><value><stats	<shotoff><value><stats	<foulcommit><value><s	<card><value><commer	<cross><value><stats><	<corner><value><stats>	<possession><value><c
<goal><value><commer	<shoton><value><stats	<shotoff><value><stats	<foulcommit><value><s	<card><value><commer	<cross><value><stats><	<corner><value><stats>	<possession><value><c
<goal><value><commer	<shoton><value><stats	<shotoff><value><stats	<foulcommit><value><s	<card><value><commer	<cross><value><stats><	<corner><value><stats>	<possession><value><c
<goal><value><commer	<shoton><value><stats	<shotoff><value><stats	<foulcommit><value><s	<card><value><commer	<cross><value><stats><	<corner><value><stats>	<possession><value><c
<goal><value><commer	<shoton><value><stats	<shotoff><value><stats	<foulcommit><value><s	<card><value><commer	<cross><value><stats><	<corner><value><stats>	<possession><value><c

Player	11060	7	id, player_api_id, player_name, player_fifa_api_id, birthday, height, weight
Player_Attributes	183978	42	id, player_fifa_api_id, player_api_id, date, overall_rating, potential, preferred_foot, attacking_work_rate, defensive_work_rate, crossing, finishing, heading_accuracy, short_passing, volleys, dribbling, curve, free_kick_accuracy, long_passing, ball_control, acceleration, sprint_speed, agility, reactions, balance, shot_power, jumping, stamina, strength, long_shots, aggression, interceptions, positioning, vision, penalties, marking, standing_tackle, sliding_tackle, gk_diving, gk_handling, gk_kicking, gk_positioning, gk_reflexes
sqlite_sequence	7	2	name, seq
Team	299	5	id, team_api_id, team_fifa_api_id, team_long_name, team_short_name
Team_Attributes	1458	25	id, team_fifa_api_id, team_api_id, date, buildUpPlaySpeed, buildUpPlaySpeedClass, buildUpPlayDribbling, buildUpPlayDribblingClass, buildUpPlayPassing, buildUpPlayPassingClass, buildUpPlayPositioningClass, chanceCreationPassing, chanceCreationPassingClass, chanceCreationCrossing, chanceCreationCrossingClass, chanceCreationShooting, chanceCreationShootingClass, chanceCreationPositioningClass, defencePressure, defencePressureClass, defenceAggression, defenceAggressionClass, defenceTeamWidth, defenceTeamWidthClass, defenceDefenderLineClass

The remaining tables (figure above) were pulled from the popular soccer game, FIFA. These included attributes for individual players as well as tactics and attributes for teams.

Datahub.io

Due to the corrupted data in the first dataset, we decided to search for more complete data online. Eventually, we found Datahub.io, which provided simplified data for the last 10 years of English Premier League games in the form of .csv files. Luckily, this dataset included some relevant statistics that were corrupted in the Kaggle dataset.

season-1819

Signup to Premium Service for additional or customised data - Get Started

Share: https://datahub.io/sports-data/eng Embed: <iframe src="https://datahub.io/sp

Date	HomeTe	AwayTe	FTHG	FTAG	FTR	HTHG	HTAG	HTR	Referee	HS	AS	HST	AST	HF	AF	HC	AC	HY	AY	HR	AR
2018-08-10	Man United	Leiceste	2	1	H	1	0	H	A Marrine	8	13	6	4	11	8	2	5	2	1	0	0
2018-08-11	Bourne	Cardiff	2	0	H	1	0	H	K Friend	12	10	4	1	11	9	7	4	1	1	0	0
2018-08-11	Fulham	Crystal Palace	0	2	A	0	1	A	M Dean	15	10	6	9	9	11	5	5	1	2	0	0
2018-08-11	Hudder	Chelsea	0	3	A	0	2	A	C Kavana	6	13	1	4	9	8	2	5	2	1	0	0
2018-08-11	Newcas	Tottenh	1	2	A	1	2	A	M Atkinso	15	15	2	5	11	12	3	5	2	2	0	0
2018-08-11	Watford	Brighto	2	0	H	1	0	H	J Moss	19	6	5	0	10	16	8	2	2	2	0	0
2018-08-11	Wolves	Everton	2	2	D	1	1	D	C Pawson	11	6	4	5	8	7	3	6	0	1	0	1
2018-08-12	Arsenal	Man City	0	2	A	0	1	A	M Oliver	9	17	3	8	11	14	2	9	2	2	0	0

Whoscored.com

In order to obtain more granular data about the events in each match, we attempted to scrape the popular football statistics website, Whoscored.com. However, due to stringent security measures put in place, we were unable to directly scrape the data. Thus we decided to manually record the values in the javascript files for the simulation.

```
import requests
from bs4 import BeautifulSoup as soup
url = "https://www.whoscored.com/Statistics"
page_html = requests.get(url)
page_soup = soup(page_html.content, 'html.parser')
```

```
from incapsula import IncapSession, RecaptchaBlocked
session = IncapSession()
try:
    response = session.get(url, bypass_crack=True)
    output = soup(response.content, 'html.parser')
except RecaptchaBlocked as e:
    raise
```

Our scraping attempts failed

R	Team	Shots pg	Shots OT pg	Dribbles pg	Fouled pg
1	Arsenal	15.1	5.6	13	11.4
2	Leicester	13.7	4.7	11.3	9.3
3	Tottenham	17.3	6.6	9.5	10.2
4	Manchester City	16.2	5.5	12	10.1
5	West Ham	14.7	4.9	10.4	11.8
6	Southampton	13.7	4.4	7.1	9.6
7	Liverpool	16.6	5.3	10.8	9.9
8	Everton	12.9	4.6	11.7	10.8
9	Chelsea	13.8	4.6	12.3	12.6
10	Manchester United	11.3	3.8	9.8	9.4
11	Watford	11.7	3.7	8.7	10.2
12	Sunderland	11.6	3.7	8	10.1
13	Crystal Palace	12.3	4	10.3	11.3
14	Swansea	11.6	3.6	7.9	11.9
15	Stoke	11	3.4	10.6	10.1
16	Newcastle United	10.4	4	9.7	9.5
17	West Bromwich Albion	10.2	2.8	7.6	10.3
18	Bournemouth	12.2	3.7	10.9	10.4
19	Aston Villa	10	2.9	11.1	9.6
20	Norwich	11	3.4	8.4	8.1

Offensive Statistics

R	Team	Shots pg	Tackles pg	Interceptions pg	Fouls pg	Offsides pg
1	Arsenal	11.8	18.6	19.4	9.2	2.3
2	Leicester	13.6	22.9	21.6	10.7	1.3
3	Tottenham	11.1	21.1	17	11.9	1.9
4	Manchester City	8.9	19.2	16.6	10.8	1.7
5	West Ham	13.7	18.7	17.5	10.1	2.2
6	Southampton	12.2	18.8	18.6	11	2.6
7	Liverpool	10.5	22.9	14.5	11.1	2.4
8	Everton	14.6	19.2	13.4	8.3	1.4
9	Chelsea	12.7	20.8	13.6	10.6	2.4
10	Manchester United	10.8	20	15.6	12.4	2
11	Watford	13.2	20.9	20.2	12.1	1.4
12	Sunderland	15	20.2	14.9	10.6	2.3
13	Crystal Palace	14.9	17.1	18.4	12.4	2.1
14	Swansea	13.7	16.8	17.9	10.3	2.2
15	Stoke	14.1	18.7	15.2	11	2.7
16	Newcastle United	14.5	20.8	16.9	11	1.6
17	West Bromwich Albion	14.4	16.9	19	10.3	1.8
18	Bournemouth	11.6	18.6	16.8	9.5	1.7
19	Aston Villa	13.2	19.2	20.2	11.3	1.1
20	Norwich	12.9	16	12.4	11	2.4

Defensive Statistics

With all the data collected, we moved on to the Data Analysis and Parameter Estimation stage of the project, primarily using Excel and Python.