Project Luther:

Scraping and Modeling Major League Soccer Data

Overview

- Introduction
- Scraping the Data
- Modeling the Data
- Challenges
- Questions

Introduction

Objective:

To explain how to use the python data stack to scrape and model Major League Soccer (MLS) goalkeeper statistics in order to predict salary.

Tools Used:

Requests
http for humans

Beautiful Soup

Pandas

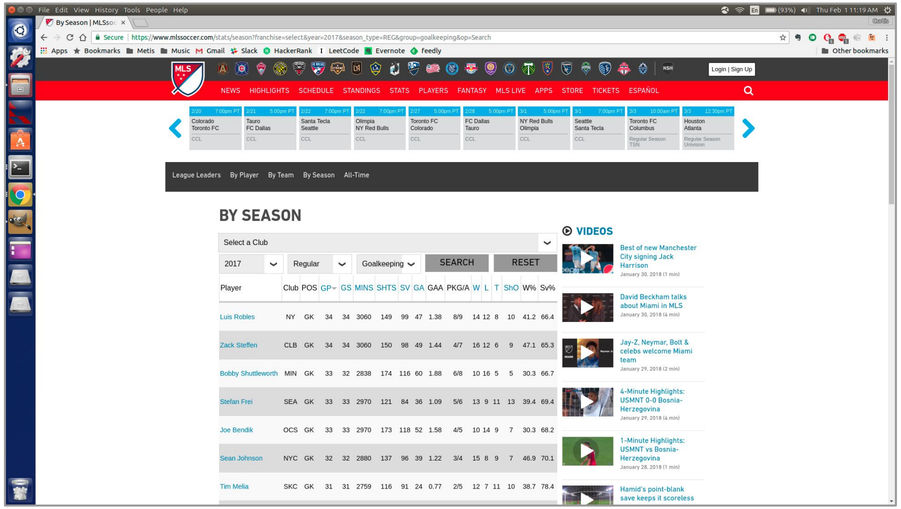
SM StatsModels
Statistics in Python

scikits learn

Scraping the Data

Data Sources

www.mlssoccer.com/stats


By Season | MLSsoc
https://www.mlssoccer.com/stats/season?franchise=select&year=2017&season_type=REG&group=goalkeeping&op=Search
NEWS HIGHLIGHTS SCHEDULE STANDINGS STATS PLAYERS FANTASY MLS LIVE APPS STORE TICKETS ESPAÑOL
Login | Sign Up
League Leaders
By Player
By Team
By Season
All-Time

BY SEASON

Select a Club

2017 | Regular | Goalkeeping | SEARCH | RESET

Player	Club	POS	GP	GS	MINS	SHTS	SV	GA	GAA	PKG/A	W	L	T	ShO	W%	Sv%
Luis Robles	NY	GK	34	34	3060	149	99	47	1.38	8/9	14	12	8	10	41.2	66.4
Zack Steffen	CLB	GK	34	34	3060	150	98	49	1.44	4/7	16	12	6	9	47.1	65.3
Bobby Shuttleworth	MIN	GK	33	32	2838	174	116	60	1.88	6/8	10	16	5	5	30.3	66.7
Stefan Frei	SEA	GK	33	33	2970	121	84	36	1.09	5/6	13	9	11	13	39.4	69.4
Joe Bendik	OCS	GK	33	33	2970	173	118	52	1.58	4/5	10	14	9	7	30.3	68.2
Sean Johnson	NYC	GK	32	32	2880	137	96	39	1.22	3/4	15	8	9	7	46.9	70.1
Tim Melia	SKC	GK	31	31	2759	116	91	24	0.77	2/5	12	7	11	10	38.7	78.4

VIDEOS

Best of new Manchester City signing Jack Harrison
January 30, 2018 (1 min)

David Beckham talks about Miami in MLS
January 30, 2018 (4 min)

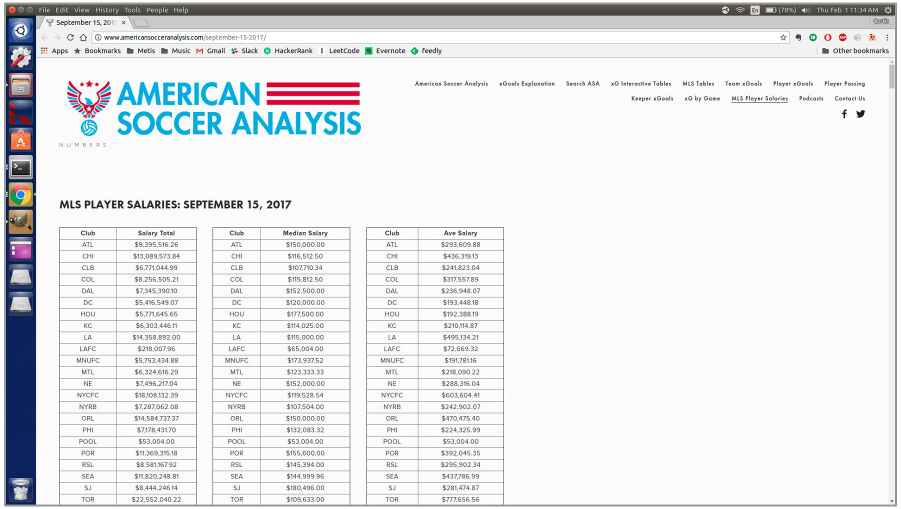
Jay-Z, Neymar, Bolt & celebs welcome Miami team
January 29, 2018 (2 min)

4-Minute Highlights: USMNT 0-0 Bosnia-Herzegovina
January 29, 2018 (4 min)

1-Minute Highlights: USMNT vs Bosnia-Herzegovina
January 28, 2018 (1 min)

Hamid's point-blank save keeps it scoreless

www.americansocceranalysis.com



MLS PLAYER SALARIES: SEPTEMBER 15, 2017

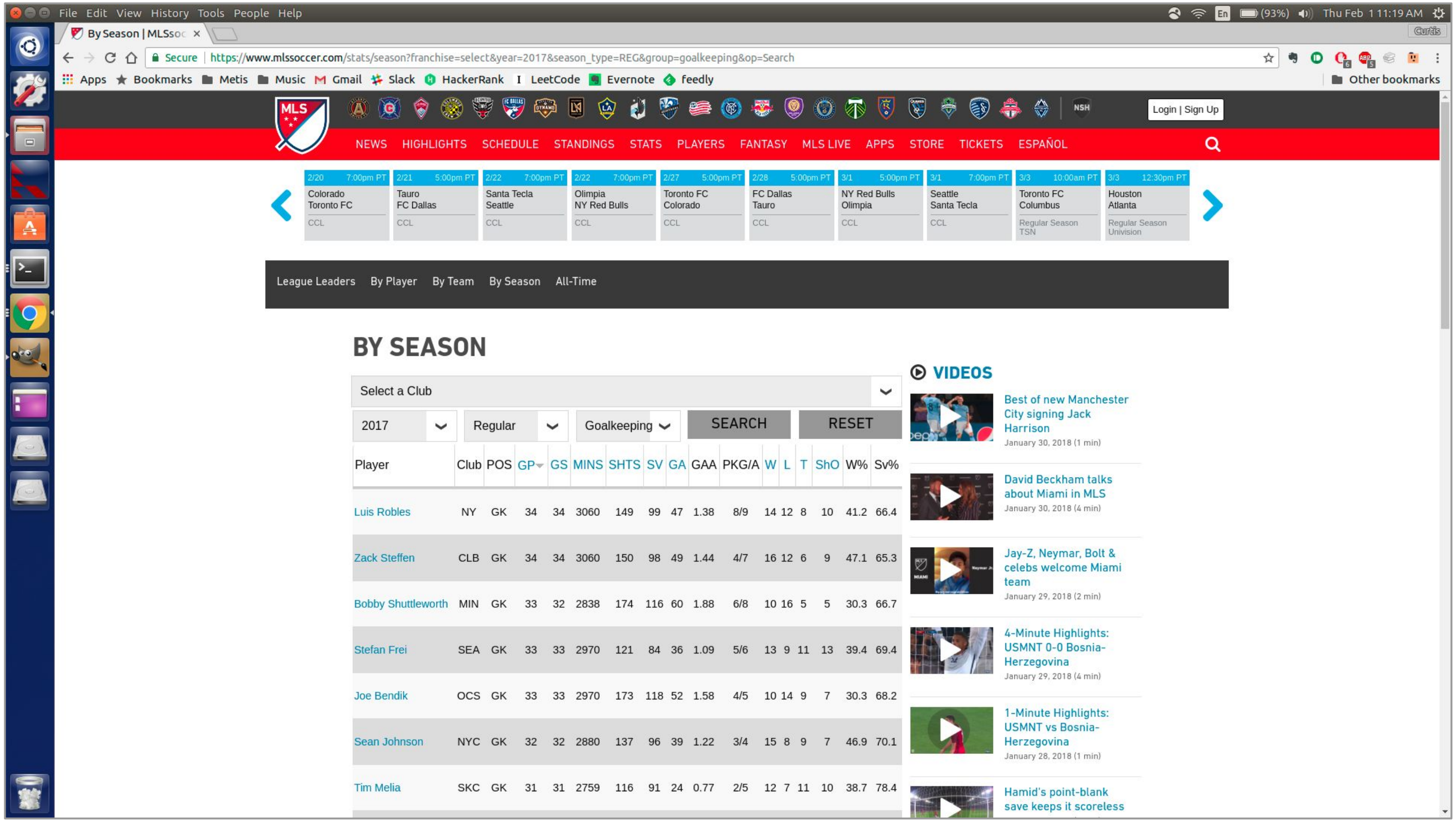
Club	Salary Total
ATL	$9,395,516.26
CHI	$13,089,573.84
CLB	$6,771,044.99
COL	$8,256,505.21
DAL	$7,345,390.10
DC	$5,416,549.07
HOU	$5,771,645.65
KC	$6,303,446.11
LA	$14,358,892.00
LAFC	$218,007.96
MNUFC	$5,753,434.88
MTL	$6,324,616.29
NE	$7,496,217.04
NYCFC	$18,108,132.39
NYRB	$7,287,062.08
ORL	$14,584,737.37
PHI	$7,178,431.70
POOL	$53,004.00
POR	$11,369,315.18
RSL	$8,581,167.92
SEA	$11,820,248.81
SJ	$8,444,246.14
TOR	$22,552,040.22

Club	Median Salary
ATL	$150,000.00
CHI	$116,512.50
CLB	$107,710.34
COL	$115,812.50
DAL	$152,500.00
DC	$120,000.00
HOU	$177,500.00
KC	$114,025.00
LA	$115,000.00
LAFC	$65,004.00
MNUFC	$173,937.52
MTL	$123,333.33
NE	$152,000.00
NYCFC	$119,528.54
NYRB	$107,504.00
ORL	$150,000.00
PHI	$132,083.32
POOL	$53,004.00
POR	$155,600.00
RSL	$145,394.00
SEA	$144,999.96
SJ	$180,496.00
TOR	$109,633.00

Club	Ave Salary
ATL	$293,609.88
CHI	$436,319.13
CLB	$241,823.04
COL	$317,557.89
DAL	$236,948.07
DC	$193,448.18
HOU	$192,388.19
KC	$210,114.87
LA	$495,134.21
LAFC	$72,669.32
MNUFC	$191,781.16
MTL	$218,090.22
NE	$288,316.04
NYCFC	$603,604.41
NYRB	$242,902.07
ORL	$470,475.40
PHI	$224,325.99
POOL	$53,004.00
POR	$392,045.35
RSL	$295,902.34
SEA	$437,786.99
SJ	$281,474.87
TOR	$777,656.56

Player statistics and salary information for the years 2007 through 2017.

What does mlssoccer.com look like?



File Edit View History Tools People Help
Thu Feb 1 11:19 AM
By Season | MLSsoc
Secure
https://www.mlssoccer.com/stats/season?franchise=select&year=2017&season_type=REG&group=goalkeeping&op=Search
Apps
Bookmarks
Metis
Music
Gmail
Slack
HackerRank
LeetCode
Evernote
feedly
Other bookmarks
MLS
Login | Sign Up
NEWS HIGHLIGHTS SCHEDULE STANDINGS STATS PLAYERS FANTASY MLS LIVE APPS STORE TICKETS ESPAÑOL

2/20 7:00pm PT	2/21 5:00pm PT	2/22 7:00pm PT	2/22 7:00pm PT	2/27 5:00pm PT	2/28 5:00pm PT	3/1 5:00pm PT	3/1 7:00pm PT	3/3 10:00am PT	3/3 12:30pm PT
Colorado Toronto FC	Tauro FC Dallas	Santa Tecla Seattle	Olimpia NY Red Bulls	Toronto FC Colorado	FC Dallas Tauro	NY Red Bulls Olimpia	Seattle Santa Tecla	Toronto FC Columbus	Houston Atlanta
CCL	CCL	CCL	CCL	CCL	CCL	CCL	CCL	Regular Season TSN	Regular Season Univision

League Leaders By Player By Team By Season All-Time

BY SEASON

Select a Club

2017 Regular Goalkeeping SEARCH RESET

Player	Club	POS	GP	GS	MINS	SHTS	SV	GA	GAA	PKG/A	W	L	T	ShO	W%	Sv%
Luis Robles	NY	GK	34	34	3060	149	99	47	1.38	8/9	14	12	8	10	41.2	66.4
Zack Steffen	CLB	GK	34	34	3060	150	98	49	1.44	4/7	16	12	6	9	47.1	65.3
Bobby Shuttleworth	MIN	GK	33	32	2838	174	116	60	1.88	6/8	10	16	5	5	30.3	66.7
Stefan Frei	SEA	GK	33	33	2970	121	84	36	1.09	5/6	13	9	11	13	39.4	69.4
Joe Bendik	OCS	GK	33	33	2970	173	118	52	1.58	4/5	10	14	9	7	30.3	68.2
Sean Johnson	NYC	GK	32	32	2880	137	96	39	1.22	3/4	15	8	9	7	46.9	70.1
Tim Melia	SKC	GK	31	31	2759	116	91	24	0.77	2/5	12	7	11	10	38.7	78.4

VIDEOS

Best of new Manchester City signing Jack Harrison
January 30, 2018 (1 min)

David Beckham talks about Miami in MLS
January 30, 2018 (4 min)

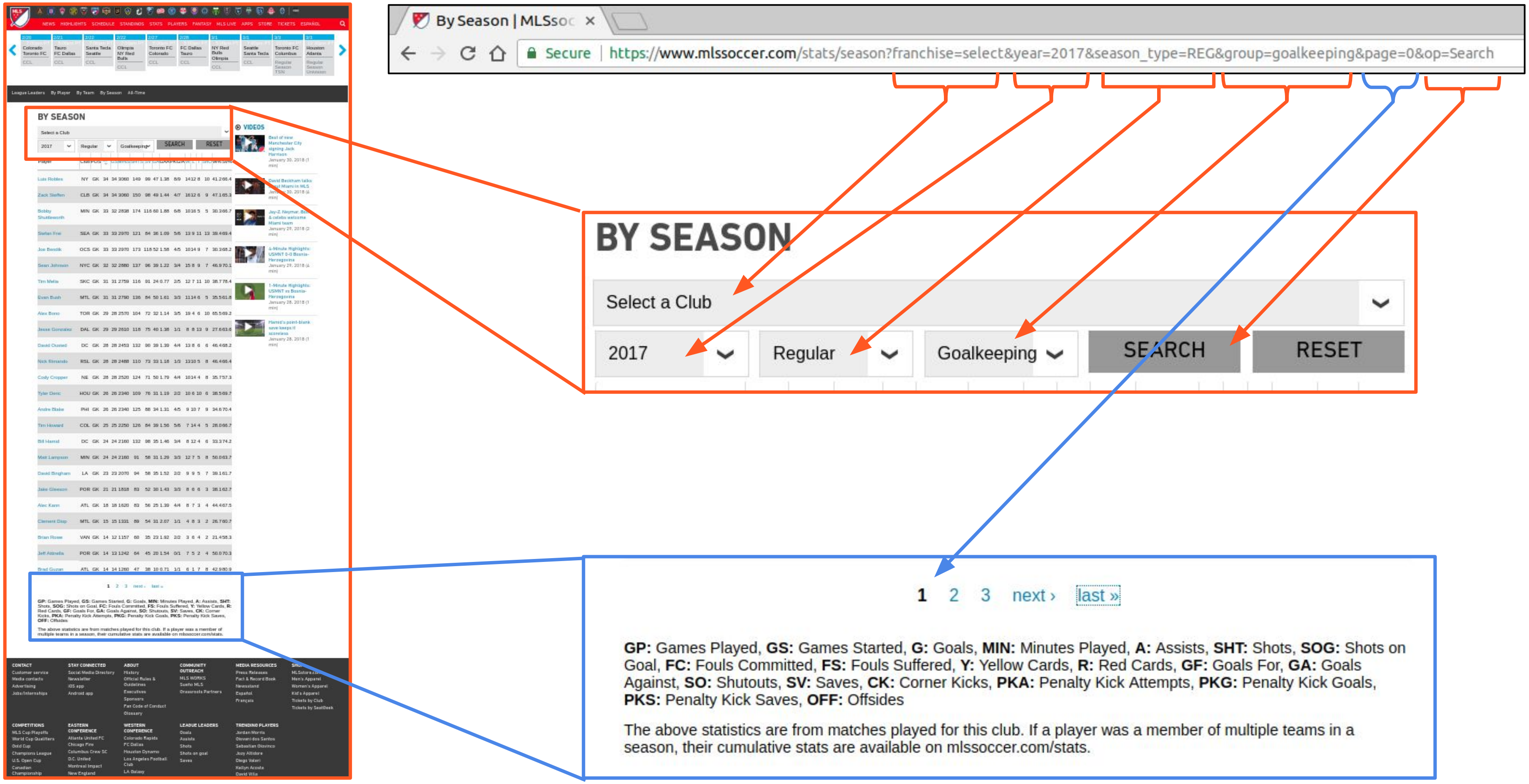
Jay-Z, Neymar, Bolt & celebs welcome Miami team
January 29, 2018 (2 min)

4-Minute Highlights: USMNT 0-0 Bosnia-Herzegovina
January 29, 2018 (4 min)

1-Minute Highlights: USMNT vs Bosnia-Herzegovina
January 28, 2018 (1 min)

Hamid's point-blank save keeps it scoreless

Identifying URL Structure



By Season | MLSsoc ×

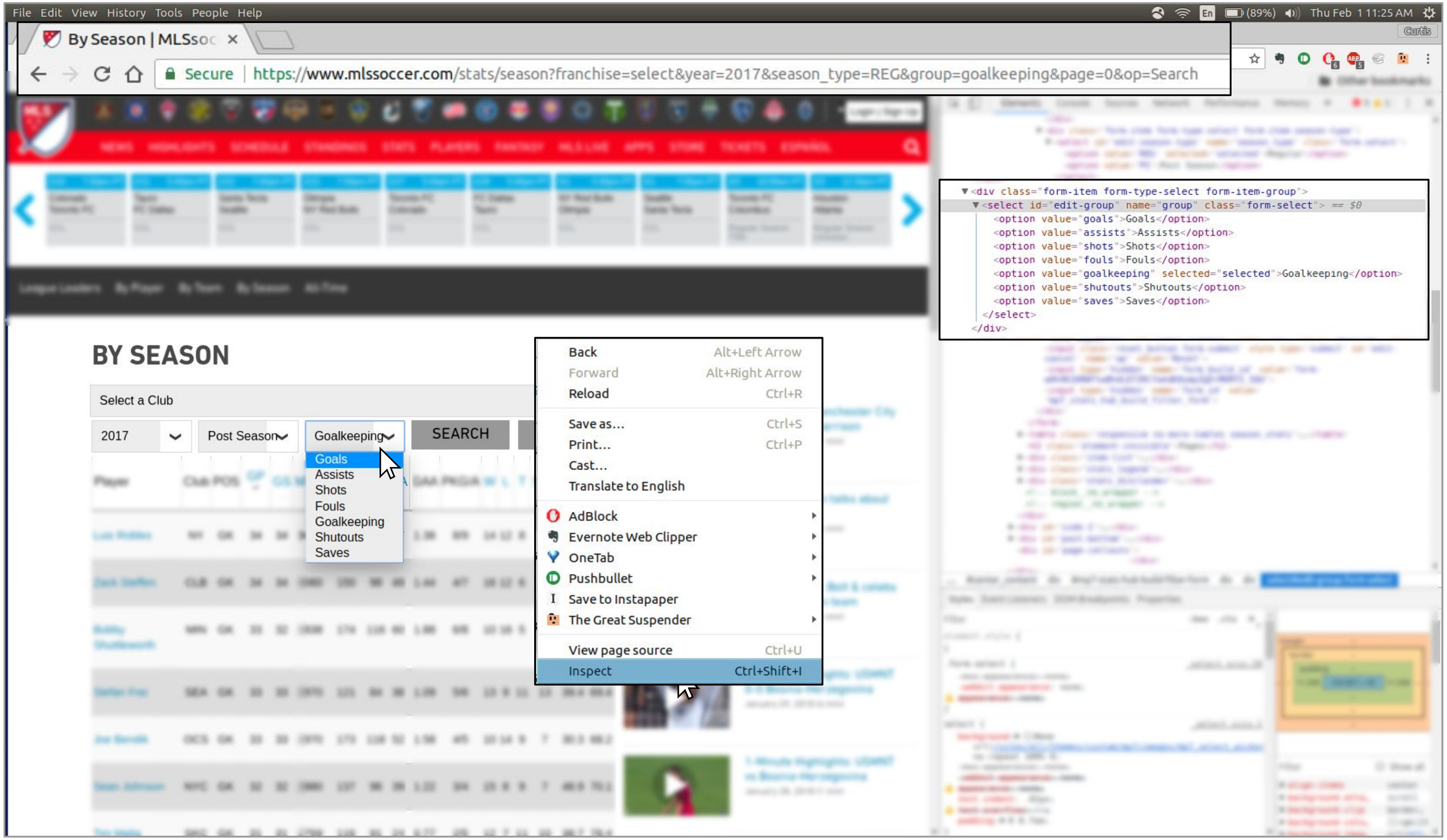
Secure | https://www.mlssoccer.com/stats/season?franchise=select&year=2017&season_type=REG&group=goalkeeping&page=0&op=Search

BY SEASON

GP: Games Played, **GS:** Games Started, **G:** Goals, **MIN:** Minutes Played, **A:** Assists, **SHT:** Shots, **SOG:** Shots on Goal, **FC:** Fouls Committed, **FS:** Fouls Suffered, **Y:** Yellow Cards, **R:** Red Cards, **GF:** Goals For, **GA:** Goals Against, **SO:** Shutouts, **SV:** Saves, **CK:** Corner Kicks, **PKA:** Penalty Kick Attempts, **PKG:** Penalty Kick Goals, **PKS:** Penalty Kick Saves, **OFF:** Offsides

The above statistics are from matches played for this club. If a player was a member of multiple teams in a season, their cumulative stats are available on mlssoccer.com/stats.

Locating Parameter Values

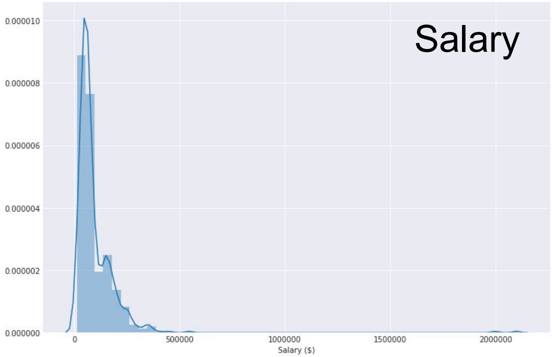


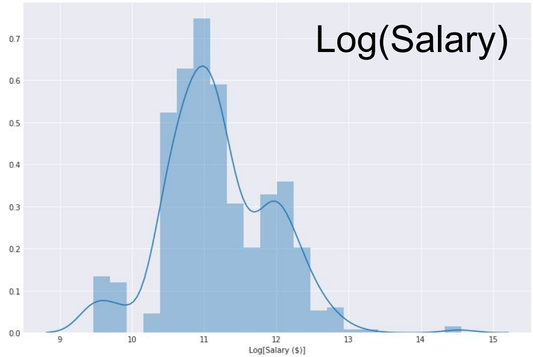
https://www.mlssoccer.com/stats/season?franchise=select&year=2017&season_type=REG&group=goalkeeping&page=0&op=Search

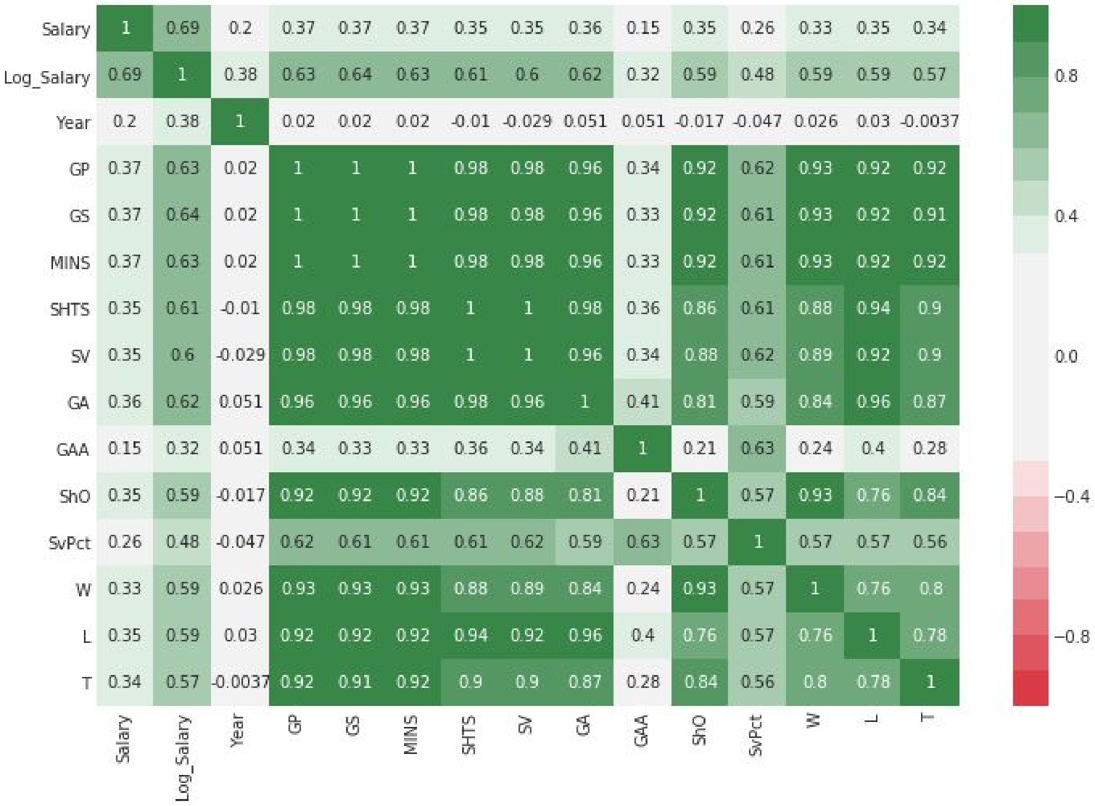
```
▼<div class="form-item form-type-select form-item-group">
  ▼<select id="edit-group" name="group" class="form-select"> == $0
      <option value="goals">Goals</option>
      <option value="assists">Assists</option>
      <option value="shots">Shots</option>
      <option value="fouls">Fouls</option>
      <option value="goalkeeping" selected="selected">Goalkeeping</option>
      <option value="shutouts">Shutouts</option>
      <option value="saves">Saves</option>
    </select>
  </div>
```

Modeling the Data

Transforming & Exploring

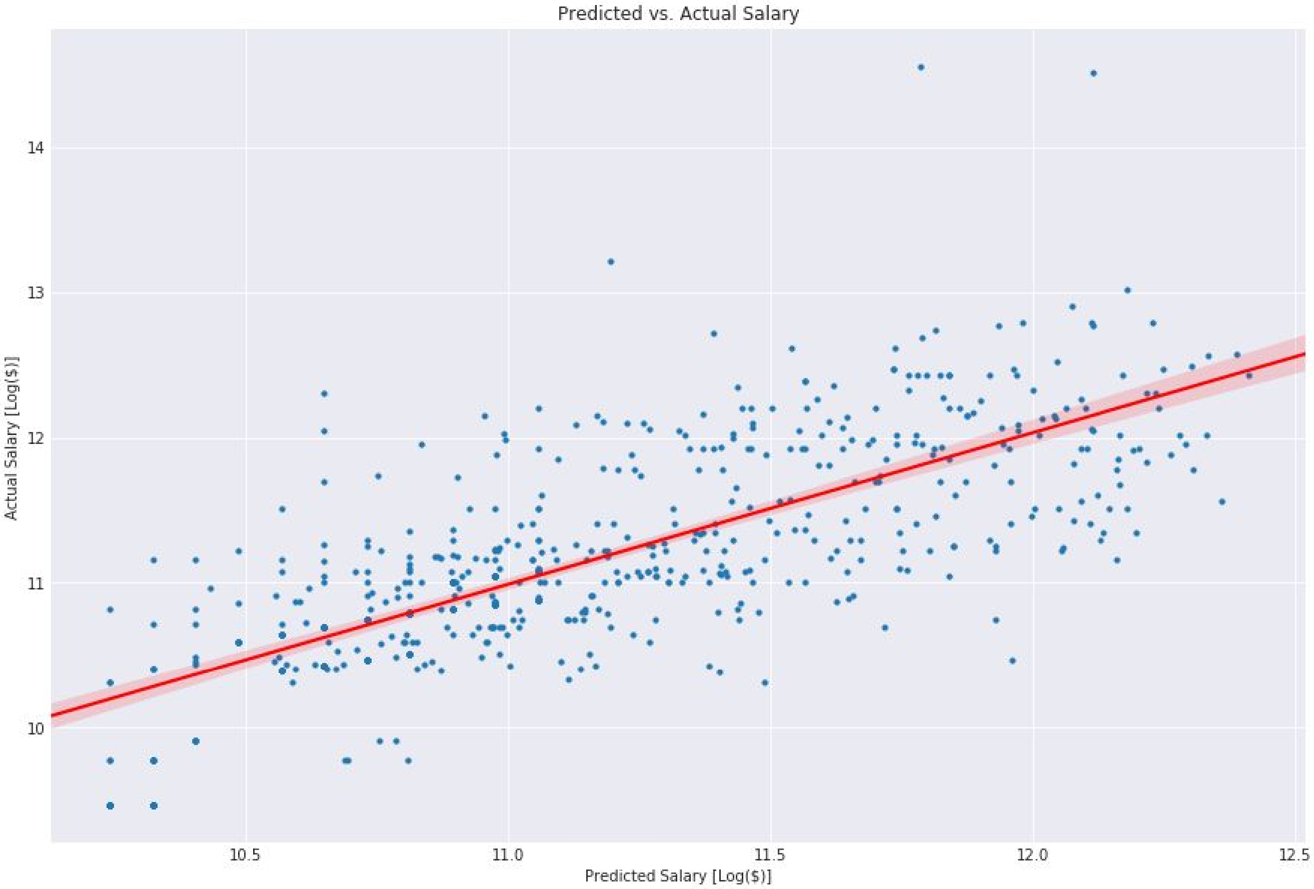




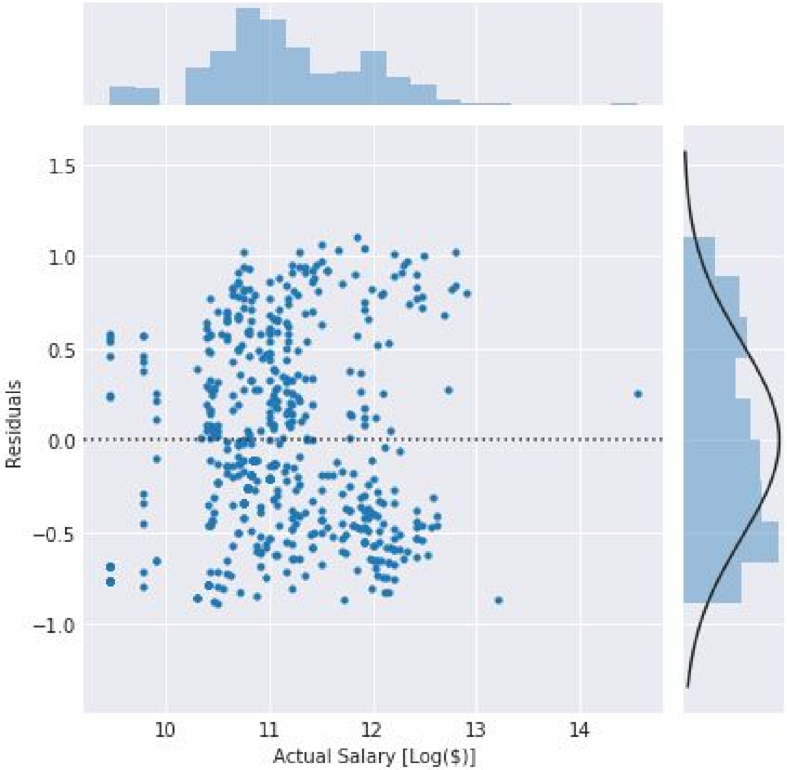


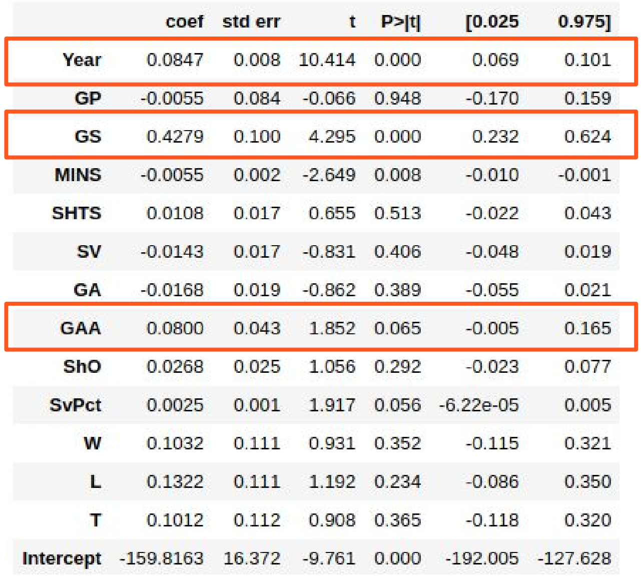
	Salary	Log_Salary	Year	GP	GS	MINS	SHTS	SV	GA	GAA	ShO	SvPct	W	L	T
Salary	1	0.69	0.2	0.37	0.37	0.37	0.35	0.35	0.36	0.15	0.35	0.26	0.33	0.35	0.34
Log_Salary	0.69	1	0.38	0.63	0.64	0.63	0.61	0.6	0.62	0.32	0.59	0.48	0.59	0.59	0.57
Year	0.2	0.38	1	0.02	0.02	0.02	-0.01	-0.029	0.051	0.051	-0.017	-0.047	0.026	0.03	-0.0037
GP	0.37	0.63	0.02	1	1	1	0.98	0.98	0.96	0.34	0.92	0.62	0.93	0.92	0.92
GS	0.37	0.64	0.02	1	1	1	0.98	0.98	0.96	0.33	0.92	0.61	0.93	0.92	0.91
MINS	0.37	0.63	0.02	1	1	1	0.98	0.98	0.96	0.33	0.92	0.61	0.93	0.92	0.92
SHTS	0.35	0.61	-0.01	0.98	0.98	0.98	1	1	0.98	0.36	0.86	0.61	0.88	0.94	0.9
SV	0.35	0.6	-0.029	0.98	0.98	0.98	1	1	0.96	0.34	0.88	0.62	0.89	0.92	0.9
GA	0.36	0.62	0.051	0.96	0.96	0.96	0.98	0.96	1	0.41	0.81	0.59	0.84	0.96	0.87
GAA	0.15	0.32	0.051	0.34	0.33	0.33	0.36	0.34	0.41	1	0.21	0.63	0.24	0.4	0.28
ShO	0.35	0.59	-0.017	0.92	0.92	0.92	0.86	0.88	0.81	0.21	1	0.57	0.93	0.76	0.84
SvPct	0.26	0.48	-0.047	0.62	0.61	0.61	0.61	0.62	0.59	0.63	0.57	1	0.57	0.57	0.56
W	0.33	0.59	0.026	0.93	0.93	0.93	0.88	0.89	0.84	0.24	0.93	0.57	1	0.76	0.8
L	0.35	0.59	0.03	0.92	0.92	0.92	0.94	0.92	0.96	0.4	0.76	0.57	0.76	1	0.78
T	0.34	0.57	-0.0037	0.92	0.91	0.92	0.9	0.9	0.87	0.28	0.84	0.56	0.8	0.78	1

Fitting a Model

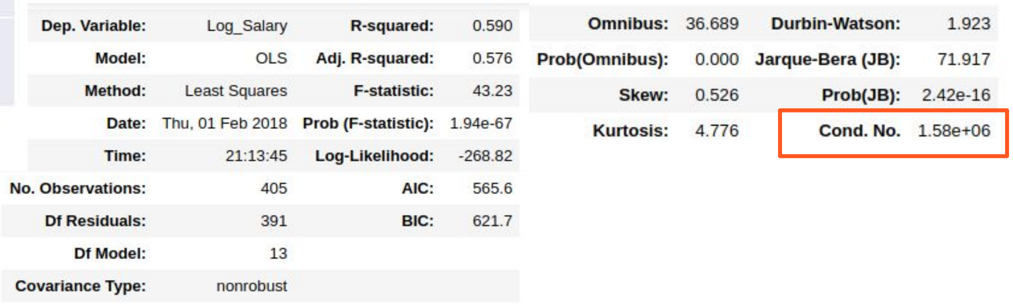


Evaluating the Fit





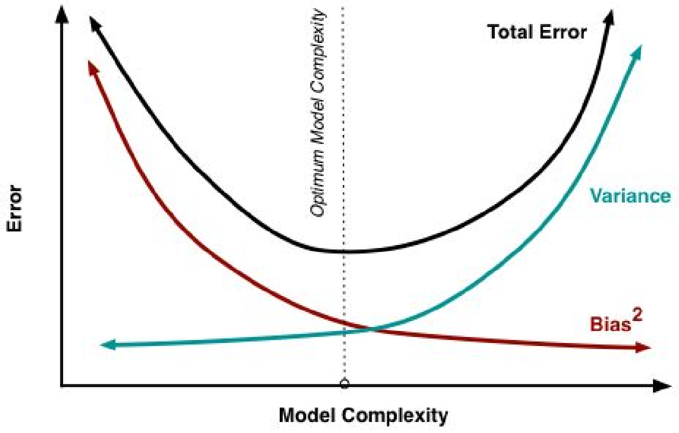
	coef	std err	t	P>|t|	[0.025	0.975]
Year	0.0847	0.008	10.414	0.000	0.069	0.101
GP	-0.0055	0.084	-0.066	0.948	-0.170	0.159
GS	0.4279	0.100	4.295	0.000	0.232	0.624
MINS	-0.0055	0.002	-2.649	0.008	-0.010	-0.001
SHTS	0.0108	0.017	0.655	0.513	-0.022	0.043
SV	-0.0143	0.017	-0.831	0.406	-0.048	0.019
GA	-0.0168	0.019	-0.862	0.389	-0.055	0.021
GAA	0.0800	0.043	1.852	0.065	-0.005	0.165
ShO	0.0268	0.025	1.056	0.292	-0.023	0.077
SvPct	0.0025	0.001	1.917	0.056	-6.22e-05	0.005
W	0.1032	0.111	0.931	0.352	-0.115	0.321
L	0.1322	0.111	1.192	0.234	-0.086	0.350
T	0.1012	0.112	0.908	0.365	-0.118	0.320
Intercept	-159.8163	16.372	-9.761	0.000	-192.005	-127.628

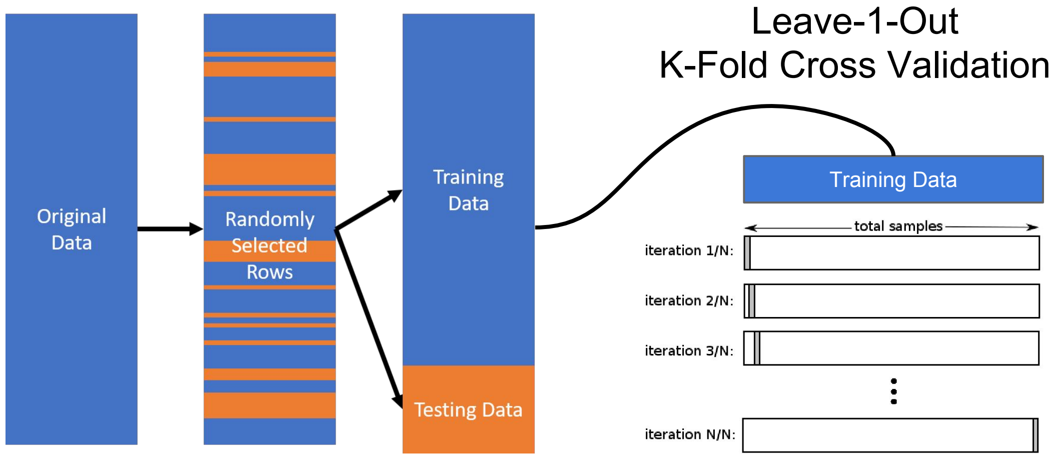


Dep. Variable:	Log_Salary	R-squared:	0.590
Model:	OLS	Adj. R-squared:	0.576
Method:	Least Squares	F-statistic:	43.23
Date:	Thu, 01 Feb 2018	Prob (F-statistic):	1.94e-67
Time:	21:13:45	Log-Likelihood:	-268.82
No. Observations:	405	AIC:	565.6
Df Residuals:	391	BIC:	621.7
Df Model:	13		
Covariance Type:	nonrobust		

Omnibus:	36.689	Durbin-Watson:	1.923
Prob(Omnibus):	0.000	Jarque-Bera (JB):	71.917
Skew:	0.526	Prob(JB):	2.42e-16
Kurtosis:	4.776	Cond. No.	1.58e+06

Generalizing the Fit





Regularized Model Scores on Test Data		
RidgeCV	LassoCV	ElasticNetCV
alpha = 0.1	alpha ~ $5.9e^{-4}$	alpha ~ $3.7e^{-4}$ (L1 = 0.975)
R^2 ~ 0.538, R^2-adj ~ 0.497	R^2 ~ 0.544, R^2-adj ~ 0.505	R^2 ~ 0.546, R^2-adj ~ 0.506

Challenges & Improvements

Issues:

- Salary is often fixed for several years at a time due to contract structure.
- Extreme outliers skew the distribution of salaries positively and the features we have cannot explain this variance.
- Many players bring intangible value to their teams that isn’t captured in their stats.
- Data integrity issues (errors in the MLS database)

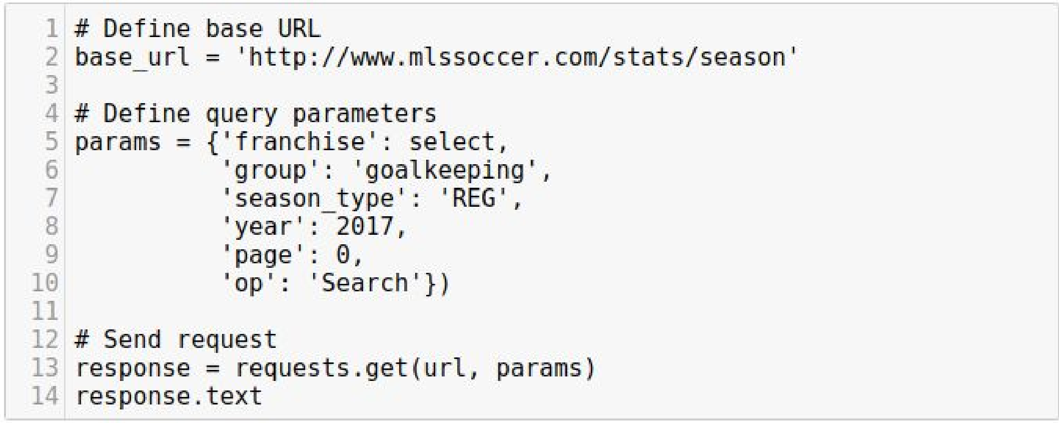
Possible Solutions:

- Add additional features to the dataset:
  - Contract signing year and any performance incentives.
  - Complete player history, not just their time in the MLS (i.e., international experience, time in other leagues)
- Data validation with other sources.

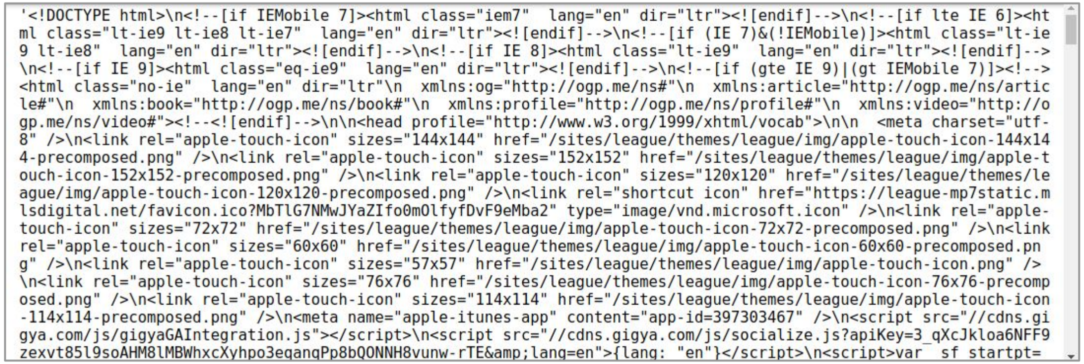
Questions?

Backup

Querying Example

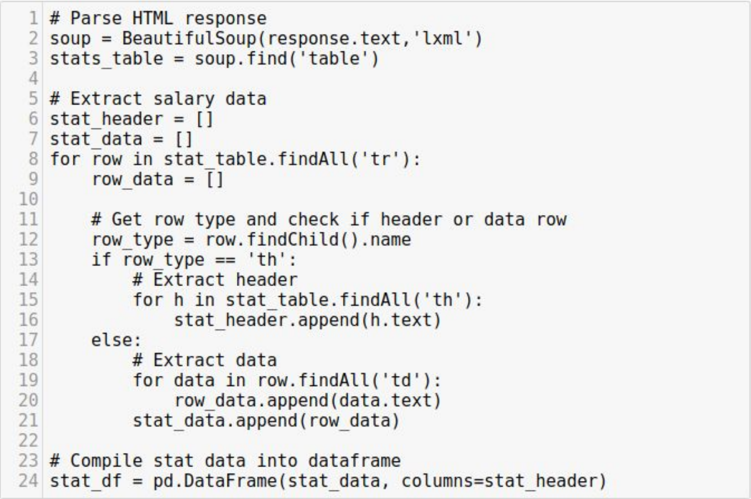


```python
# Define base URL
base_url = 'http://www.mlssoccer.com/stats/season'

# Define query parameters
params = {'franchise': select,
          'group': 'goalkeeping',
          'season_type': 'REG',
          'year': 2017,
          'page': 0,
          'op': 'Search'})

# Send request
response = requests.get(url, params)
response.text
```

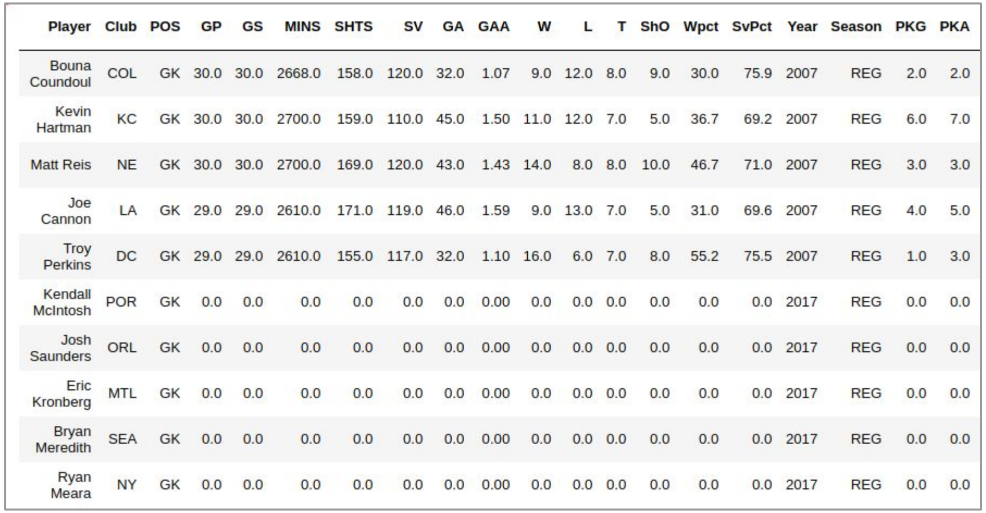


```
'<!DOCTYPE html>\n<!--[if IEMobile 7]><html class="iem7"  lang="en" dir="ltr"><![endif]-->\n<!--[if lte IE 6]><ht
ml class="lt-ie9 lt-ie8 lt-ie7"  lang="en" dir="ltr"><![endif]-->\n<!--[if (IE 7)&(!IEMobile)]><html class="lt-ie
9 lt-ie8"  lang="en" dir="ltr"><![endif]-->\n<!--[if IE 8]><html class="lt-ie9"  lang="en" dir="ltr"><![endif]-->
\n<!--[if IE 9]><html class="eq-ie9"  lang="en" dir="ltr"><![endif]-->\n<!--[if (gte IE 9)|(gt IEMobile 7)]><!-->
<html class="no-ie"  lang="en" dir="ltr"\n  xmlns:og="http://ogp.me/ns#"\n  xmlns:article="http://ogp.me/ns/artic
le#"\n  xmlns:book="http://ogp.me/ns/book#"\n  xmlns:profile="http://ogp.me/ns/profile#"\n  xmlns:video="http://o
gp.me/ns/video#"><!--<![endif]-->\n\n<head profile="http://www.w3.org/1999/xhtml/vocab">\n\n  <meta charset="utf-
8" />\n<link rel="apple-touch-icon" sizes="144x144" href="/sites/league/themes/league/img/apple-touch-icon-144x14
4-precomposed.png" />\n<link rel="apple-touch-icon" sizes="152x152" href="/sites/league/themes/league/img/apple-t
ouch-icon-152x152-precomposed.png" />\n<link rel="apple-touch-icon" sizes="120x120" href="/sites/league/themes/le
ague/img/apple-touch-icon-120x120-precomposed.png" />\n<link rel="shortcut icon" href="https://league-mp7static.m
lsdigital.net/favicon.ico?MbTlG7NMwJYaZIfo0mOlfyfDvF9eMba2" type="image/vnd.microsoft.icon" />\n<link rel="apple-
touch-icon" sizes="72x72" href="/sites/league/themes/league/img/apple-touch-icon-72x72-precomposed.png" />\n<link
rel="apple-touch-icon" sizes="60x60" href="/sites/league/themes/league/img/apple-touch-icon-60x60-precomposed.pn
g" />\n<link rel="apple-touch-icon" sizes="57x57" href="/sites/league/themes/league/img/apple-touch-icon.png" />
\n<link rel="apple-touch-icon" sizes="76x76" href="/sites/league/themes/league/img/apple-touch-icon-76x76-precomp
osed.png" />\n<link rel="apple-touch-icon" sizes="114x114" href="/sites/league/themes/league/img/apple-touch-icon
-114x114-precomposed.png" />\n<meta name="apple-itunes-app" content="app-id=397303467" />\n<script src="//cdns.gi
gya.com/js/gigyaGAIntegration.js"></script>\n<script src="//cdns.gigya.com/js/socialize.js?apiKey=3_qXcJkloa6NFF9
zexvt85l9soAHM8lMBWhxcXyhpo3eqanqPp8bQONNH8vunw-rTE&amp;lang=en">{lang: "en"}</script>\n<script>var _sf_startpt=
```

Parsing and Compiling



```python
# Parse HTML response
soup = BeautifulSoup(response.text,'lxml')
stats_table = soup.find('table')

# Extract salary data
stat_header = []
stat_data = []
for row in stat_table.findAll('tr'):
    row_data = []

    # Get row type and check if header or data row
    row_type = row.findChild().name
    if row_type == 'th':
        # Extract header
        for h in stat_table.findAll('th'):
            stat_header.append(h.text)
    else:
        # Extract data
        for data in row.findAll('td'):
            row_data.append(data.text)
        stat_data.append(row_data)

# Compile stat data into dataframe
stat_df = pd.DataFrame(stat_data, columns=stat_header)
```



Player	Club	POS	GP	GS	MINS	SHTS	SV	GA	GAA	W	L	T	ShO	Wpct	SvPct	Year	Season	PKG	PKA
Bouna Coundoul	COL	GK	30.0	30.0	2668.0	158.0	120.0	32.0	1.07	9.0	12.0	8.0	9.0	30.0	75.9	2007	REG	2.0	2.0
Kevin Hartman	KC	GK	30.0	30.0	2700.0	159.0	110.0	45.0	1.50	11.0	12.0	7.0	5.0	36.7	69.2	2007	REG	6.0	7.0
Matt Reis	NE	GK	30.0	30.0	2700.0	169.0	120.0	43.0	1.43	14.0	8.0	8.0	10.0	46.7	71.0	2007	REG	3.0	3.0
Joe Cannon	LA	GK	29.0	29.0	2610.0	171.0	119.0	46.0	1.59	9.0	13.0	7.0	5.0	31.0	69.6	2007	REG	4.0	5.0
Troy Perkins	DC	GK	29.0	29.0	2610.0	155.0	117.0	32.0	1.10	16.0	6.0	7.0	8.0	55.2	75.5	2007	REG	1.0	3.0
Kendall McIntosh	POR	GK	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	2017	REG	0.0	0.0
Josh Saunders	ORL	GK	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	2017	REG	0.0	0.0
Eric Kronberg	MTL	GK	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	2017	REG	0.0	0.0
Bryan Meredith	SEA	GK	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	2017	REG	0.0	0.0
Ryan Meara	NY	GK	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	2017	REG	0.0	0.0

Linking Datasets