
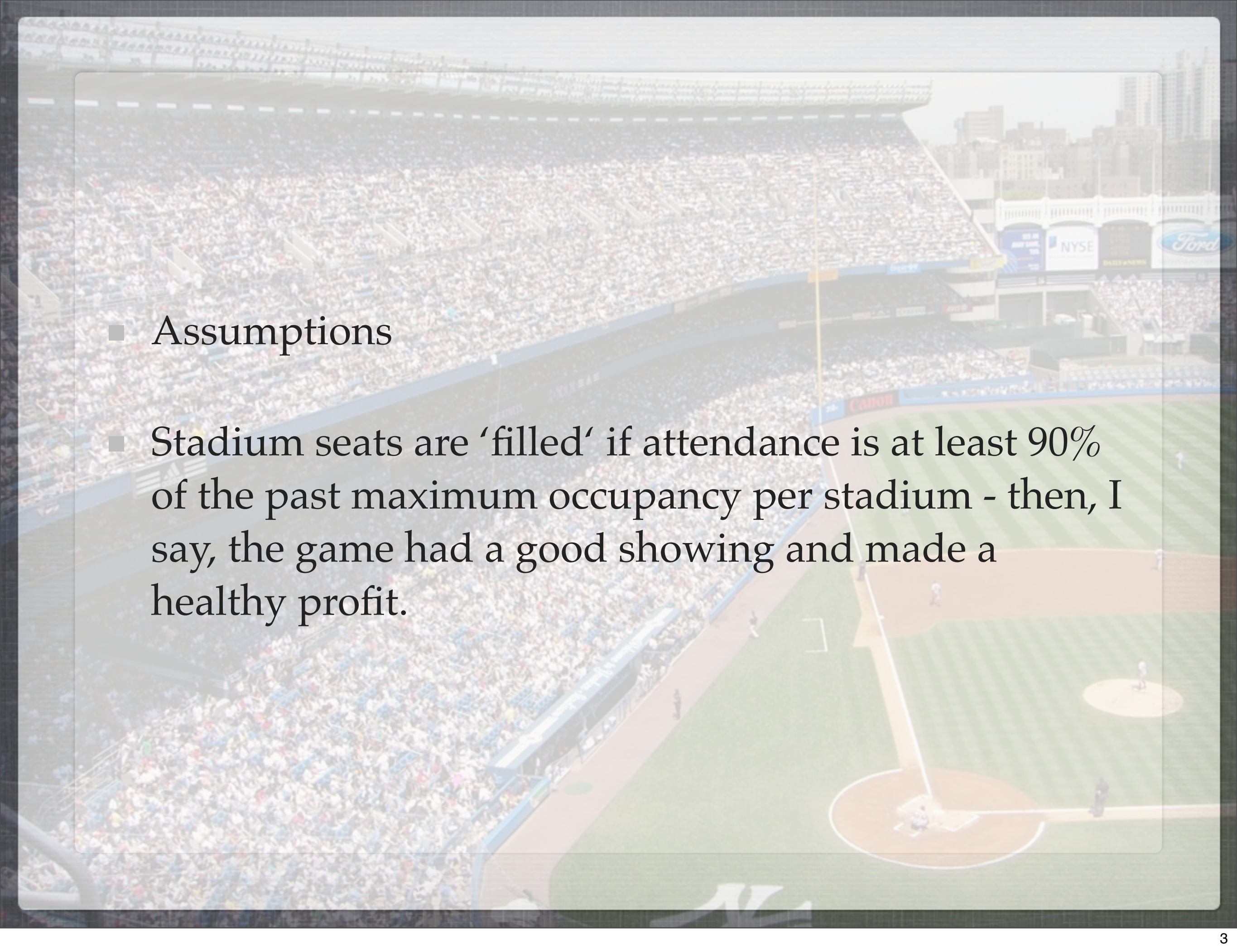


BASEBALL GAME ATTENDANCE PREDICTOR

- 
- An aerial photograph of a large baseball stadium, likely Yankee Stadium, filled with a massive crowd. The field is visible in the lower right, with players on the bases and pitcher's mound. The outfield is lined with various advertisements, including NYSE and Ford. The stadium's architecture and surrounding city skyline are visible in the background.
- From a marketing perspective, I took a look at regular season baseball game data from 2000 - 2012 to predict if a stadium reached 'capacity' seats sold, and thus made a healthy profit, in year 2013.

- 
- An aerial view of a large baseball stadium filled with spectators. The field is visible in the lower right, with players in white uniforms. The stands are packed with fans, and various advertisements like 'NYSE' and 'Ford' are visible on the outfield fence. The sky is overcast.
- Assumptions
 - Stadium seats are ‘filled’ if attendance is at least 90% of the past maximum occupancy per stadium - then, I say, the game had a good showing and made a healthy profit.

DATA SOURCE



- Retrosheet.org is awesome.
- The information used here was obtained free of charge from and is copyrighted by Retrosheet. Interested parties may contact Retrosheet at "www.retrosheet.org".
- Regular season game data:

	Date	Day	Visitors	VGameNumber	Home	HGameNumber	VScore	HScore	DayNight	ParkID	Attendance	VHRs	HHRs
0	20000329	Wed	CHN	1	NYN	1	5	3	N	TOK01	55000	2	1
1	20000330	Thu	NYN	2	CHN	2	5	1	N	TOK01	55000	1	0
2	20000403	Mon	COL	1	ATL	1	0	2	D	ATL02	42255	0	2
3	20000403	Mon	MIL	1	CIN	1	3	3	D	CIN08	55596	0	1
4	20000403	Mon	SFN	1	FLO	1	4	6	N	MIA01	35101	1	0
5	20000403	Mon	L AN	1	MON	1	10	4	N	MON02	51249	2	2

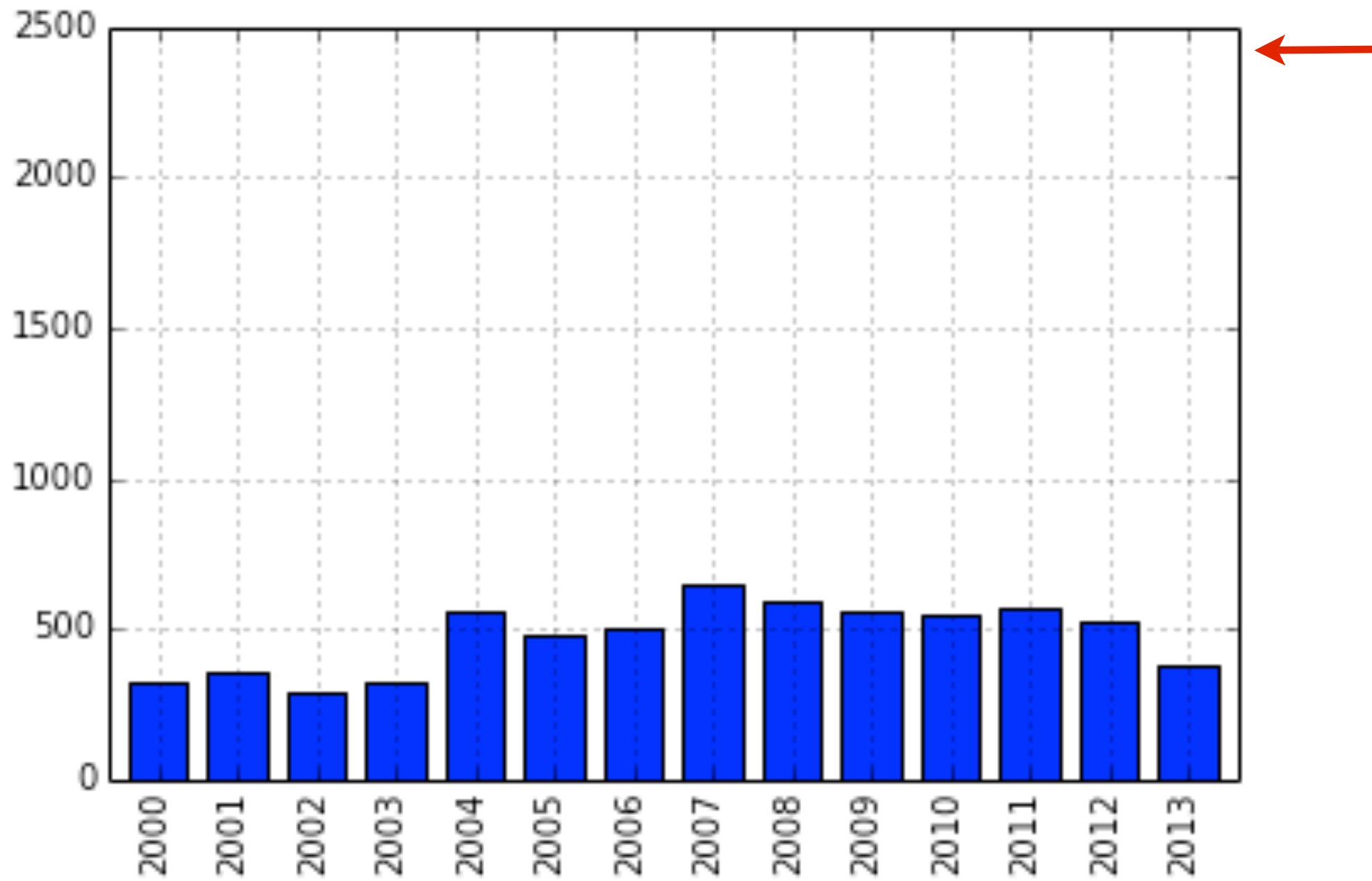
DERIVATION OF ISPROFIT COLUMN

- `bddf['PercAttn'] = [bddf.Attendance[i] / (bddf['Attendance'].groupby(bddf.ParkID==bddf.ParkID[i]).max()) for i in range(0,len(bddf))]`

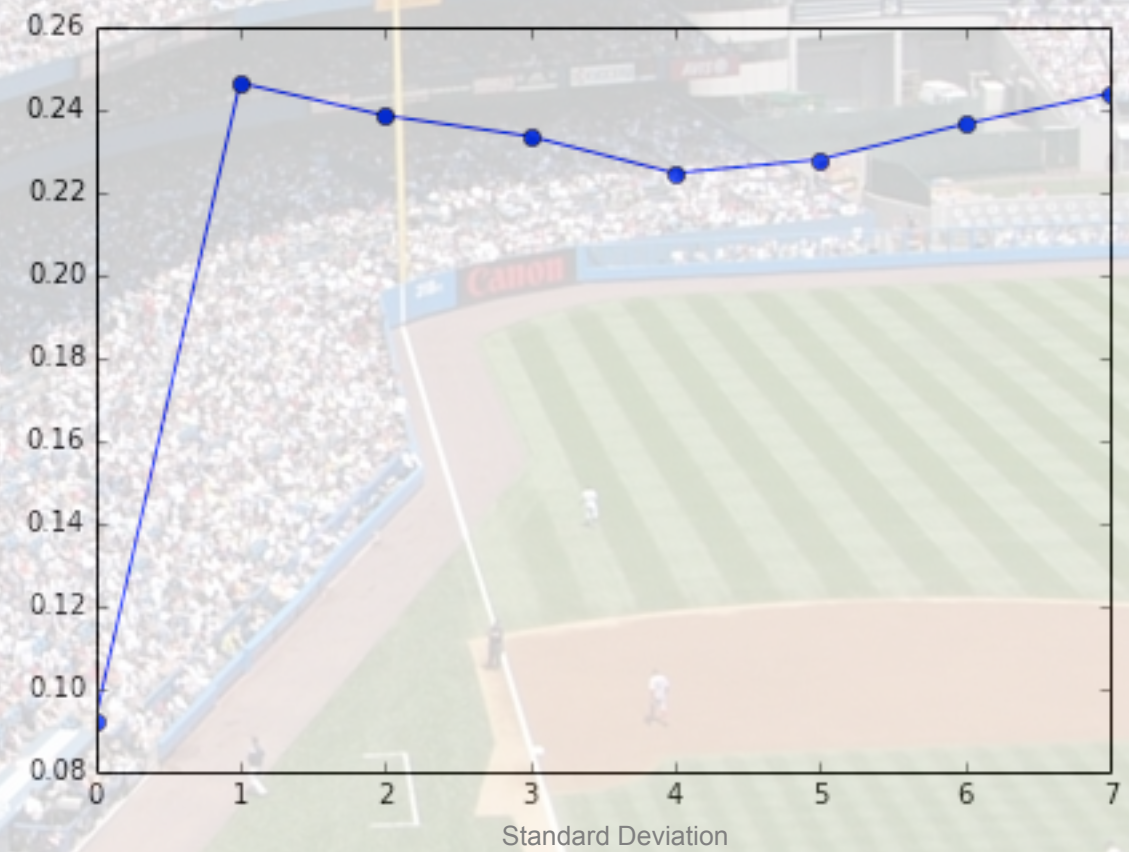
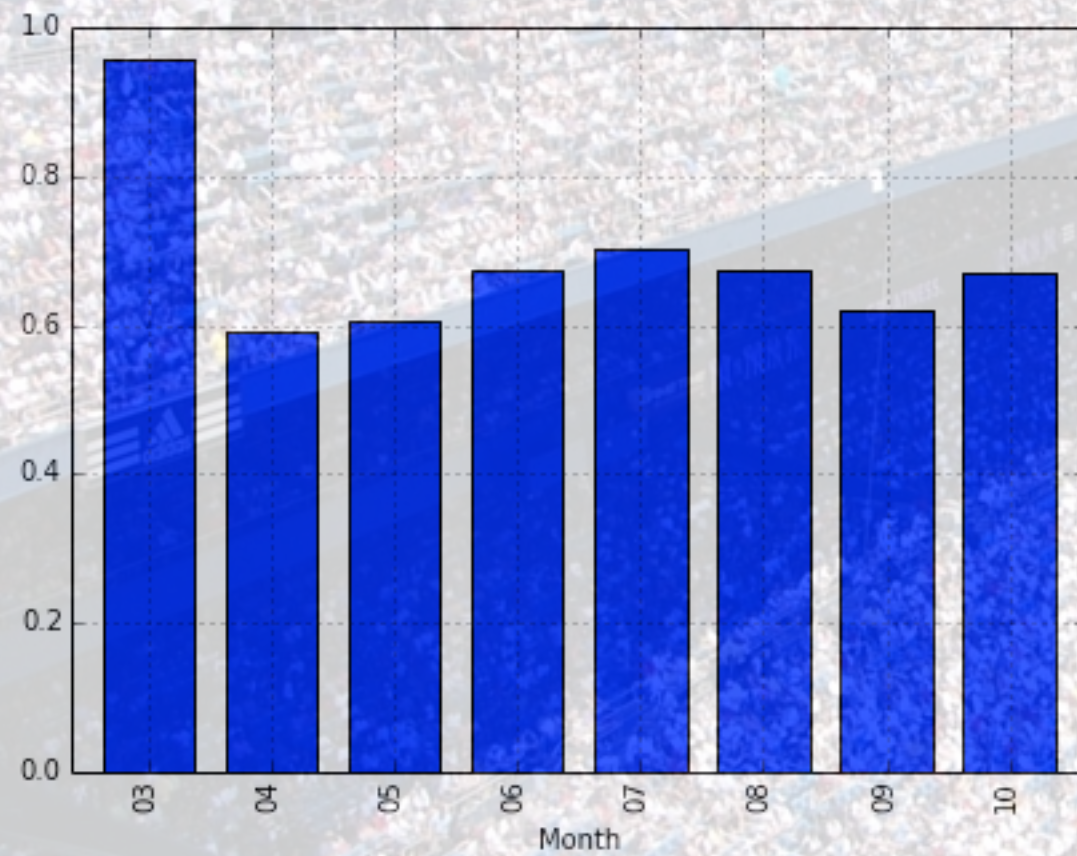
```
ParkID
False    61707
True     38540
dtype: float64
```

- `bddf['PercAttn'] = [float(str(bddf['PercAttn'][i]).split()[4]) for i in range(0,len(bddf))]`
- `bddf['IsProfit'] = [1 if bddf.PercAttn[i]>.9 else 0 for i in range(0,len(bddf))]`

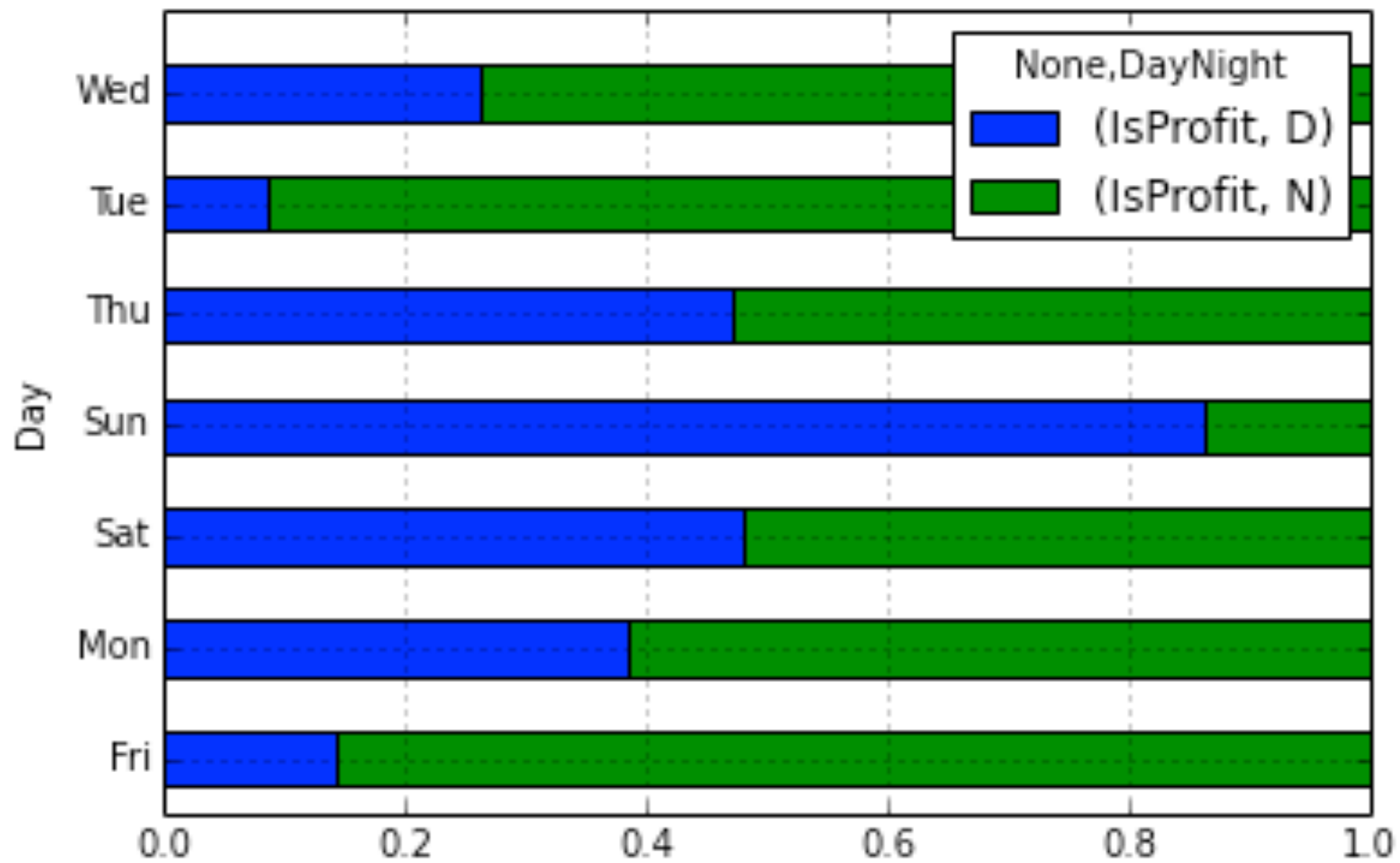
COUNT OF PROFITABLE GAMES BY YEAR



MEDIAN PERCENT ATTENDANCE BY MONTH



DAY OR NIGHT GAMES BY DAY OF WEEK



DATA SHAPE FOR LOGISTIC REGRESSION CLASSIFIER

	IsProfit	Year	NightGame	Fri	Mon	Sat	Sun	Thu	Tue	Wed	03	04	05	06	07	08	09	10
0	1	2000	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
1	1	2000	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0
2	0	2000	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
3	1	2000	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
4	0	2000	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0

↑
Predicting

- Year is not used in classifier and is there only to split out the 2013 as the test data to match the prediction.

Accuracy Score on Training Data = 80.08%

```
(array(['NightGame', 'Fri', 'Mon', 'Sat', 'Sun', 'Thu', 'Tue', 'Wed', '03',  
      '04', '05', '06', '07', '08', '09', '10'], dtype=object),  
array([[ -0.55662725,  0.3426472, -0.07197398,  0.54976186, -0.1833822 ,  
        -0.43499443, -0.36703733, -0.48627613,  1.76993844, -0.61750127,  
        -0.53565465, -0.14351234, -0.07644627, -0.27662125, -0.55035451,  
        -0.22110317]]))
```


TESTING AGAINST 2013 DATA

```
In [50]: did_it_work = [1 if predicted_Y[i] == test_set_Y[i+31580] else 0 for i in range(0, len(predicted_Y))]
```

```
In [51]: did_it_work.count(1), len(did_it_work), (did_it_work.count(1)*1.00 / len(did_it_work) * 100)
```

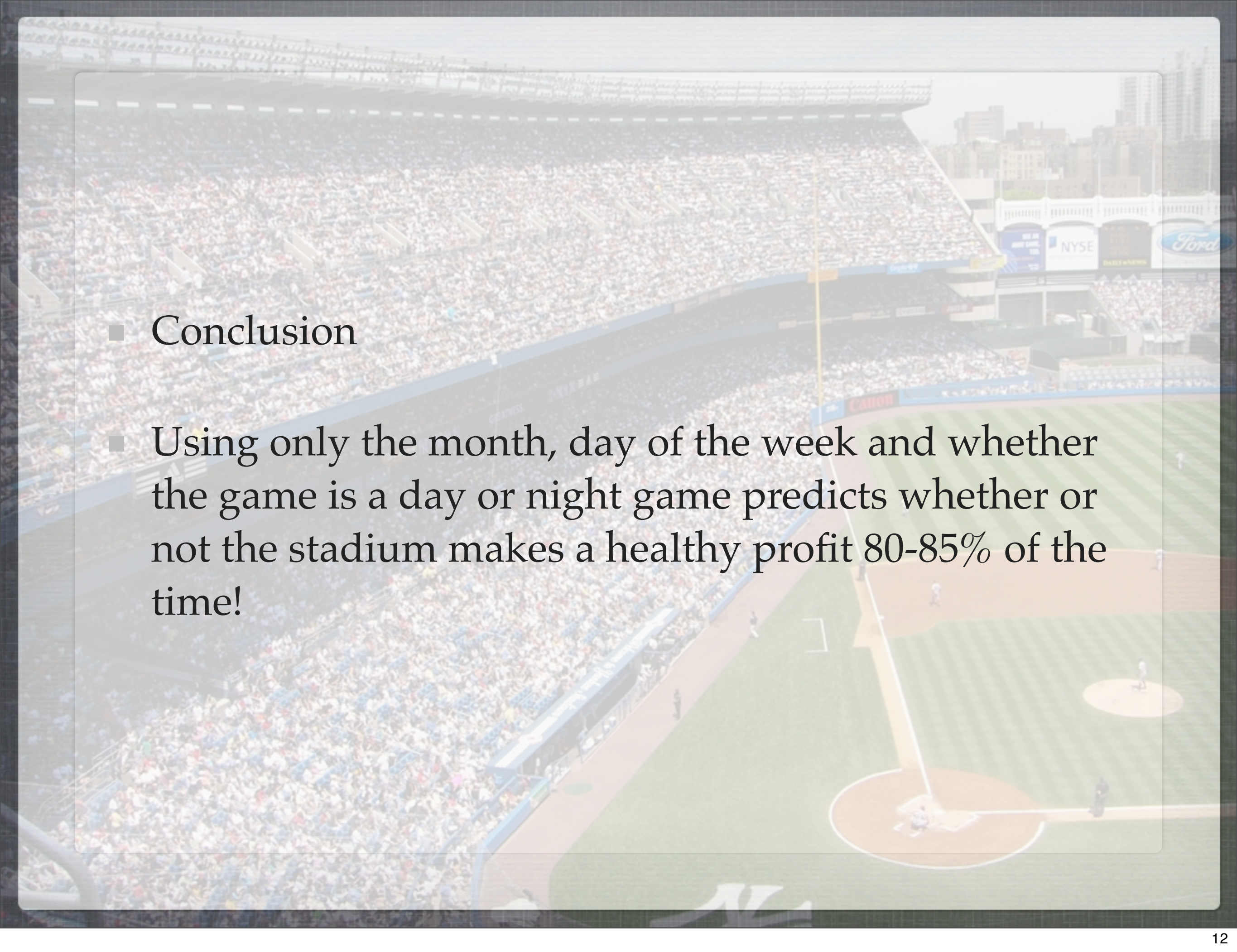
```
Out[51]: (2048, 2431, 84.24516659810777)
```

VS.

```
In [53]: clf.score(test_set_X, test_set_Y)
```

```
Out[53]: 0.84245166598107779
```



- 
- A wide-angle, high-altitude photograph of a large baseball stadium, likely Yankee Stadium, filled with a massive crowd of spectators. The field is visible in the lower right, with players positioned on the bases and pitcher's mound. The outfield is lined with various advertisements, including NYSE and Ford. The sky is overcast.
- Conclusion
 - Using only the month, day of the week and whether the game is a day or night game predicts whether or not the stadium makes a healthy profit 80-85% of the time!