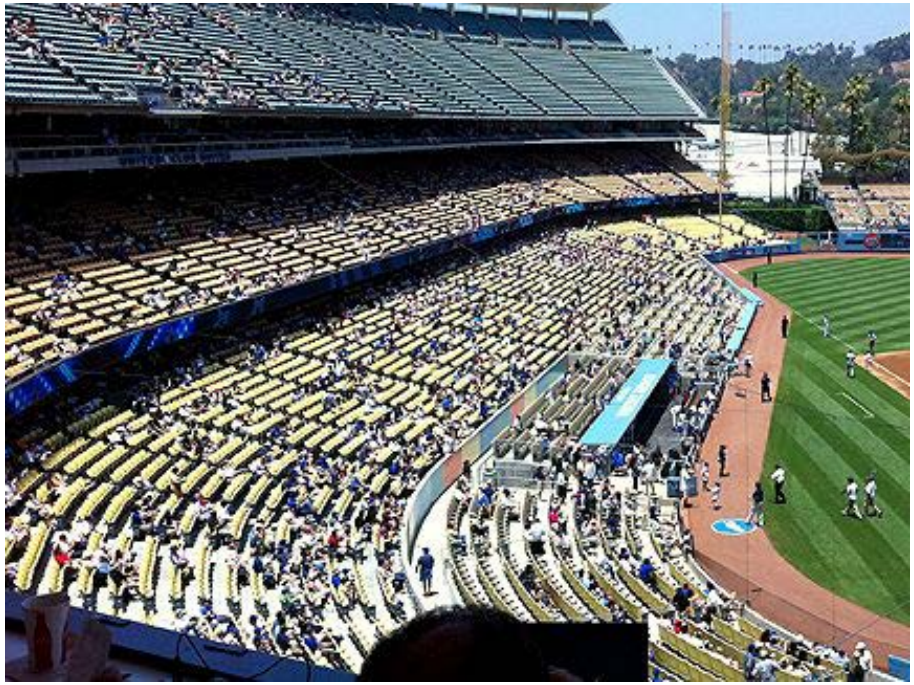
# Scenario Introduction

School of Information Studies
Syracuse University

# Welcome to the World of Analytics

- Obtain
- Scrub
- **Explore**
- **Model**
- **iNterpret**



# Our Challenge This Week?

# How Do We Attract Fans?


# But Which Promotion Is Best?

| month | day | attend | opponent | Temp | skies | day_night | cap | shirt | fireworks | bobble head |
|---|---|---|---|---|---|---|---|---|---|---|
| APR | 10 | 56000 | Pirates | 67 | Clear | Day | NO | NO | NO | NO |
| APR | 11 | 29729 | Pirates | 58 | Cloudy | Night | NO | NO | NO | NO |
| APR | 12 | 28328 | Pirates | 57 | Cloudy | Night | NO | NO | NO | NO |
| APR | 13 | 31601 | Padres | 54 | Cloudy | Night | NO | NO | YES | NO |
| APR | 14 | 46549 | Padres | 57 | Cloudy | Night | NO | NO | NO | NO |
| APR | 15 | 38359 | Padres | 65 | Clear | Day | NO | NO | NO | NO |
| APR | 23 | 26376 | Braves | 60 | Cloudy | Night | NO | NO | NO | NO |
| APR | 24 | 44014 | Braves | 63 | Cloudy | Night | NO | NO | NO | NO |




School of Information Studies
Syracuse University

# Data Review

School of Information Studies
Syracuse University

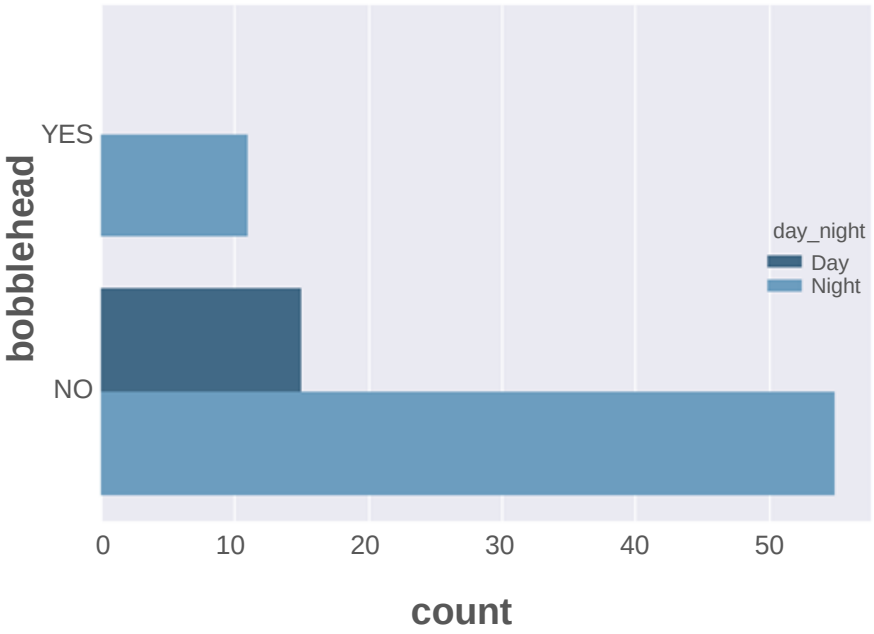
# Attendance Data

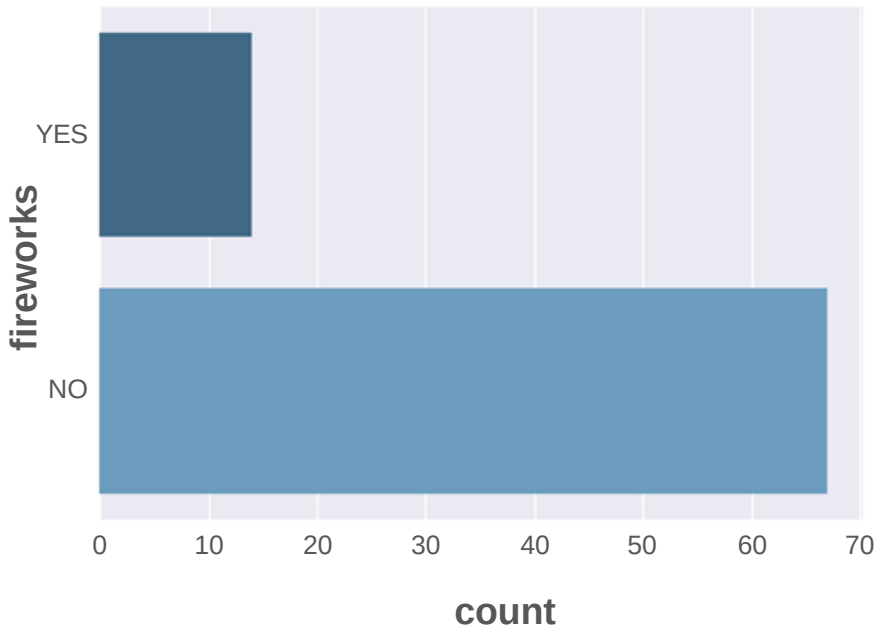
| month | day | attend | day_of_week | opponent | temp | skies | day_night | cap | shirt | fireworks | bobblehead |
|---|---|---|---|---|---|---|---|---|---|---|---|
| APR | 10 | 56000 | Tuesday | Pirates | 67 | Clear | Day | NO | NO | NO | NO |
| APR | 11 | 29729 | Wednesday | Pirates | 58 | Cloudy | Night | NO | NO | NO | NO |
| APR | 12 | 28328 | Thursday | Pirates | 57 | Cloudy | Night | NO | NO | NO | NO |
| APR | 13 | 31601 | Friday | Padres | 54 | Cloudy | Night | NO | NO | YES | NO |
| APR | 14 | 46549 | Saturday | Padres | 57 | Cloudy | Night | NO | NO | NO | NO |
| APR | 15 | 38359 | Sunday | Padres | 65 | Clear | Day | NO | NO | NO | NO |
| APR | 23 | 26376 | Monday | Braves | 60 | Cloudy | Night | NO | NO | NO | NO |
| APR | 24 | 44014 | Tuesday | Braves | 63 | Cloudy | Night | NO | NO | NO | NO |
| APR | 25 | 26345 | Wednesday | Braves | 64 | Cloudy | Night | NO | NO | NO | NO |
| APR | 27 | 44807 | Friday | Nationals | 66 | Clear | Night | NO | NO | YES | NO |
| APR | 28 | 54242 | Saturday | Nationals | 71 | Clear | Night | NO | NO | NO | YES |
| APR | 29 | 48753 | Sunday | Nationals | 74 | Clear | Day | NO | YES | NO | NO |
| MAY | 7 | 43713 | Monday | Giants | 67 | Clear | Night | NO | NO | NO | NO |
| MAY | 8 | 32799 | Tuesday | Giants | 75 | Clear | Night | NO | NO | NO | NO |
| MAY | 9 | 33993 | Wednesday | Giants | 71 | Clear | Night | NO | NO | NO | NO |
| MAY | 11 | 35591 | Friday | Rockies | 65 | Clear | Night | NO | NO | YES | NO |
| MAY | 12 | 33735 | Saturday | Rockies | 65 | Clear | Night | NO | NO | NO | NO |
| MAY | 13 | 49124 | Sunday | Rockies | 70 | Clear | Day | NO | NO | NO | NO |
| MAY | 14 | 24312 | Monday | Snakes | 67 | Clear | Night | NO | NO | NO | NO |
| MAY | 15 | 47077 | Tuesday | Snakes | 70 | Clear | Night | NO | NO | NO | YES |
| MAY | 18 | 40906 | Friday | Cardinals | 64 | Clear | Night | NO | NO | YES | NO |
| MAY | 19 | 39383 | Saturday | Cardinals | 67 | Clear | Night | NO | NO | NO | NO |
| MAY | 20 | 44005 | Sunday | Cardinals | 77 | Clear | Night | NO | NO | NO | NO |
| MAY | 25 | 36283 | Friday | Astros | 59 | Cloudy | Night | NO | NO | YES | NO |

# Summary Statistics

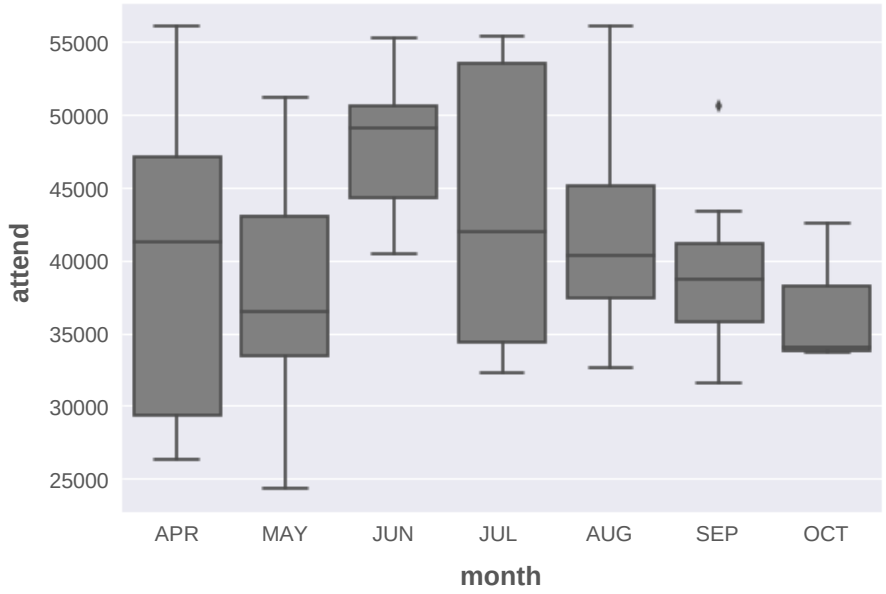
- Low – 24,312
- High – 56,000
- Mean – 41,040
- Number of promotions: 30
  - Ball cap – 2
  - Shirt – 3
  - Fireworks – 14
  - Bobblehead – 11

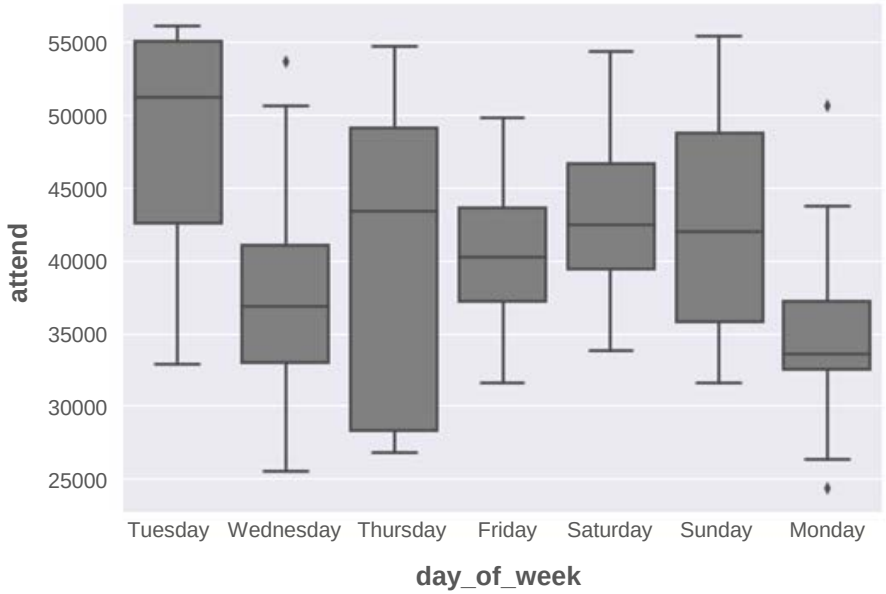
# Promotion by Type



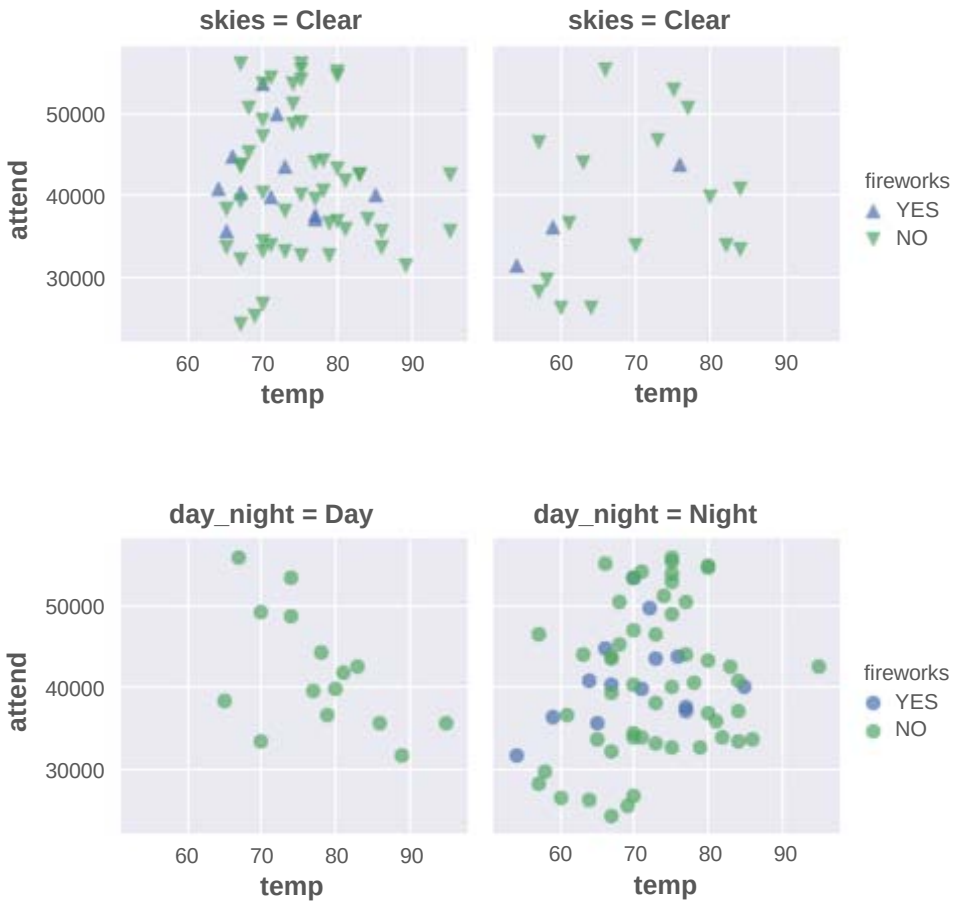



# Promotion by Time

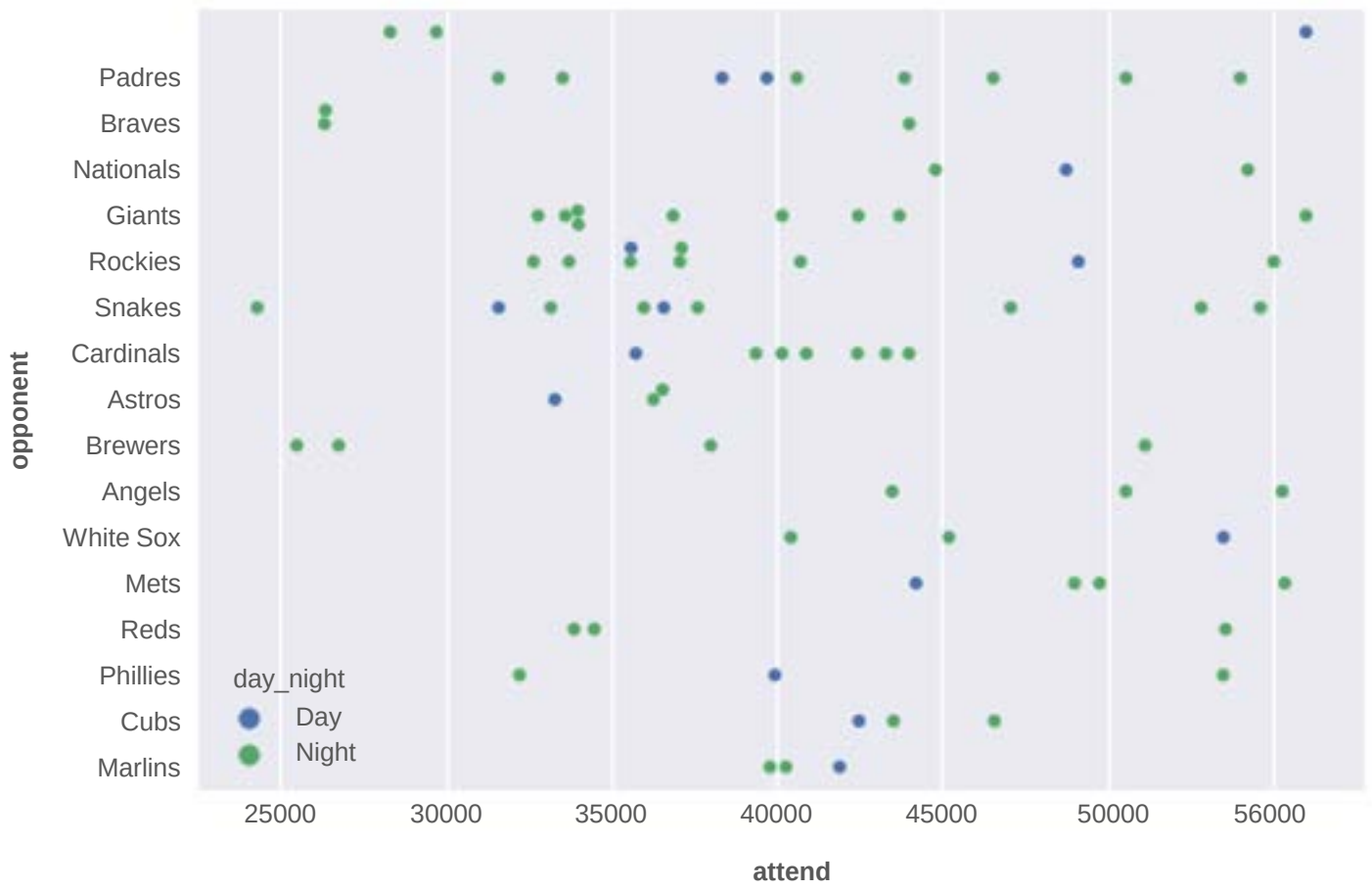


# Lattice Plots



# Attendance by Team



# Recommendation

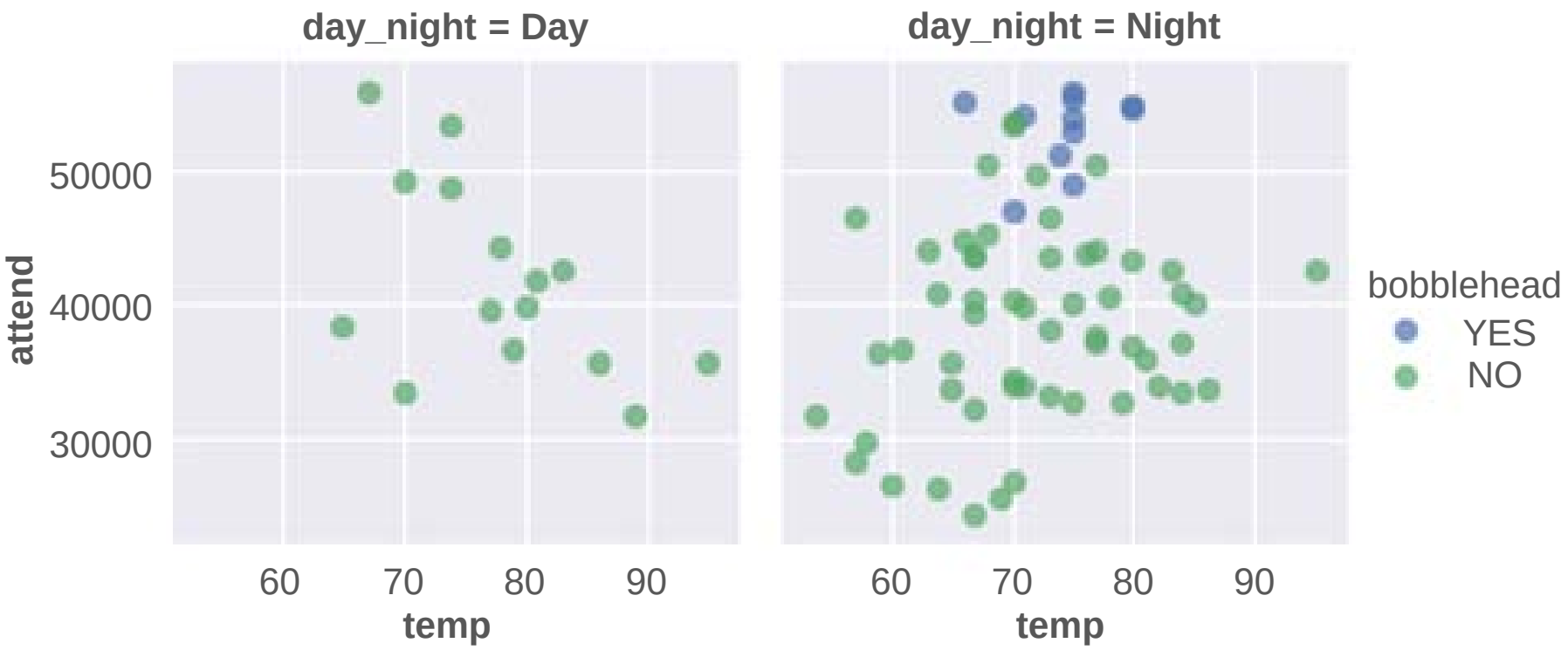
School of Information Studies
Syracuse University

# Our Challenge Was …

# How Do We Attract Fans?



# Which Promotion Is Best?



# Recommendation

- Estimated Effect of Bobblehead Promotion:
  - An extra 10,715 fans in stands
  - An extra $340,000 in ticket revenue
  - An extra $390,000 in revenue during game

# Describing

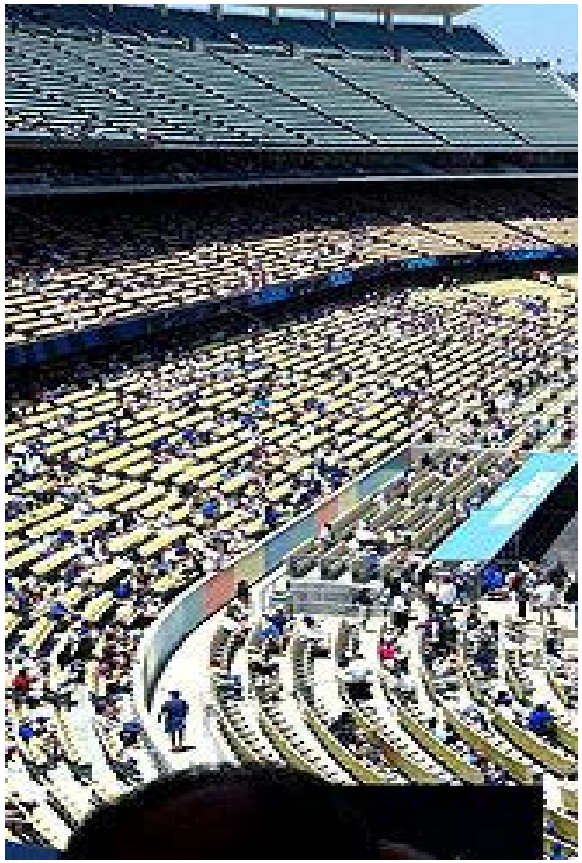
School of Information Studies
Syracuse University

# Attendance Data

| month | day | attend | day_of_week | opponent | temp | skies | day_night | cap | shirt | fireworks | bobblehead |
|---|---|---|---|---|---|---|---|---|---|---|---|
| APR | 10 | 56000 | Tuesday | Pirates | 67 | Clear | Day | NO | NO | NO | NO |
| APR | 11 | 29729 | Wednesday | Pirates | 58 | Cloudy | Night | NO | NO | NO | NO |
| APR | 12 | 28328 | Thursday | Pirates | 57 | Cloudy | Night | NO | NO | NO | NO |
| APR | 13 | 31601 | Friday | Padres | 54 | Cloudy | Night | NO | NO | YES | NO |
| APR | 14 | 46549 | Saturday | Padres | 57 | Cloudy | Night | NO | NO | NO | NO |
| APR | 15 | 38359 | Sunday | Padres | 65 | Clear | Day | NO | NO | NO | NO |
| APR | 23 | 26376 | Monday | Braves | 60 | Cloudy | Night | NO | NO | NO | NO |
| APR | 24 | 44014 | Tuesday | Braves | 63 | Cloudy | Night | NO | NO | NO | NO |
| APR | 25 | 26345 | Wednesday | Braves | 64 | Cloudy | Night | NO | NO | NO | NO |
| APR | 27 | 44807 | Friday | Nationals | 66 | Clear | Night | NO | NO | YES | NO |
| APR | 28 | 54242 | Saturday | Nationals | 71 | Clear | Night | NO | NO | NO | YES |
| APR | 29 | 48753 | Sunday | Nationals | 74 | Clear | Day | NO | YES | NO | NO |
| MAY | 7 | 43713 | Monday | Giants | 67 | Clear | Night | NO | NO | NO | NO |
| MAY | 8 | 32799 | Tuesday | Giants | 75 | Clear | Night | NO | NO | NO | NO |
| MAY | 9 | 33993 | Wednesday | Giants | 71 | Clear | Night | NO | NO | NO | NO |
| MAY | 11 | 35591 | Friday | Rockies | 65 | Clear | Night | NO | NO | YES | NO |
| MAY | 12 | 33735 | Saturday | Rockies | 65 | Clear | Night | NO | NO | NO | NO |
| MAY | 13 | 49124 | Sunday | Rockies | 70 | Clear | Day | NO | NO | NO | NO |
| MAY | 14 | 24312 | Monday | Snakes | 67 | Clear | Night | NO | NO | NO | NO |
| MAY | 15 | 47077 | Tuesday | Snakes | 70 | Clear | Night | NO | NO | NO | YES |
| MAY | 18 | 40906 | Friday | Cardinals | 64 | Clear | Night | NO | NO | YES | NO |
| MAY | 19 | 39383 | Saturday | Cardinals | 67 | Clear | Night | NO | NO | NO | NO |
| MAY | 20 | 44005 | Sunday | Cardinals | 77 | Clear | Night | NO | NO | NO | NO |
| MAY | 25 | 36283 | Friday | Astros | 59 | Cloudy | Night | NO | NO | YES | NO |

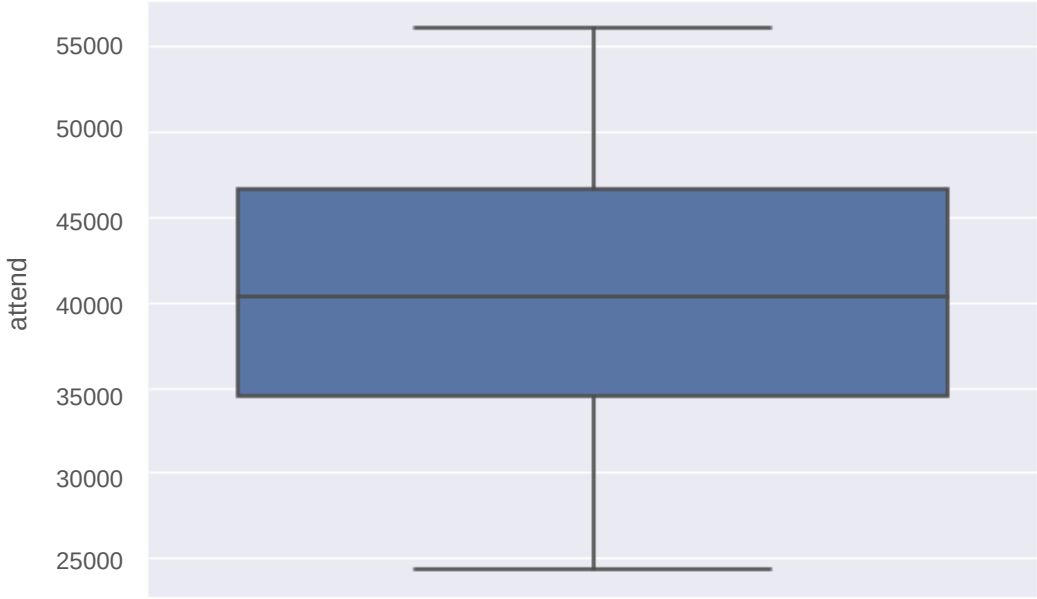
# Summary Statistics

- Low – 24,312
- High – 56,000
- Mean – 41,040
- Number of promotions: 30
  - Ball cap – 2
  - Shirt – 3
  - Fireworks – 14
  - Bobblehead – 11



# Summary Statistics

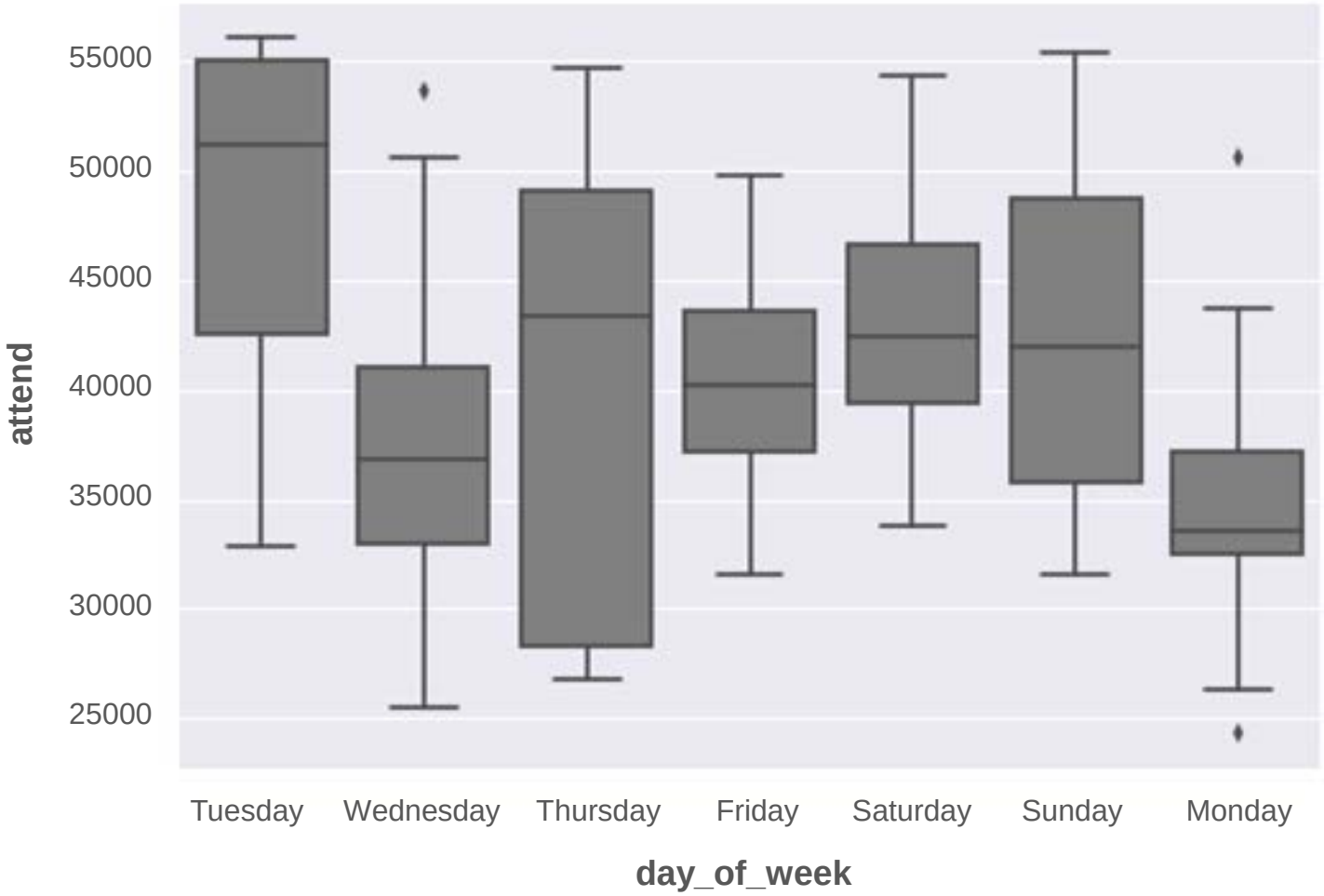
- Counts
- Maximum
- Minimum
- Quartiles
- Mean/media/mode
- Standard deviation



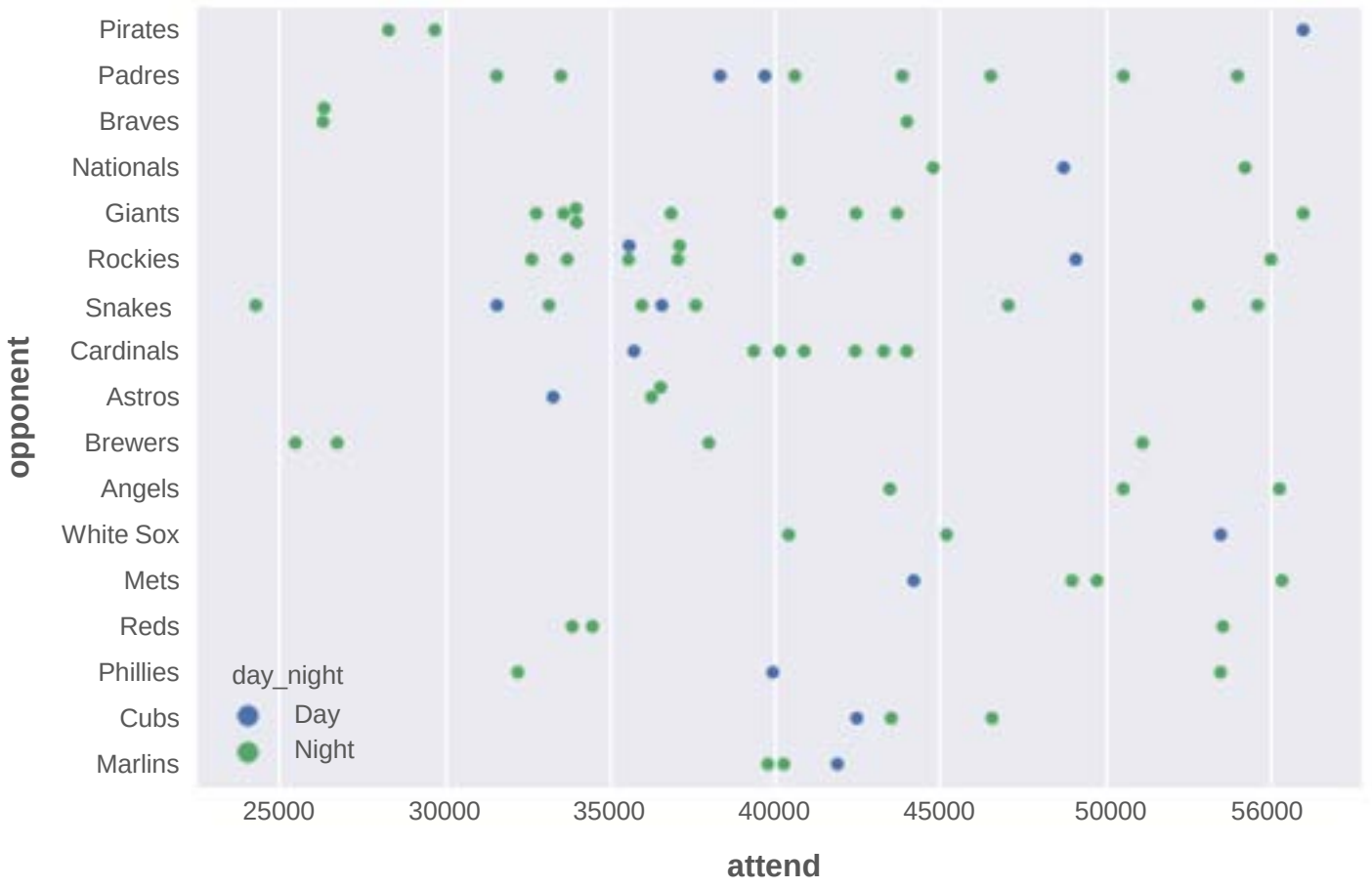
# Describing

School of Information Studies
Syracuse University

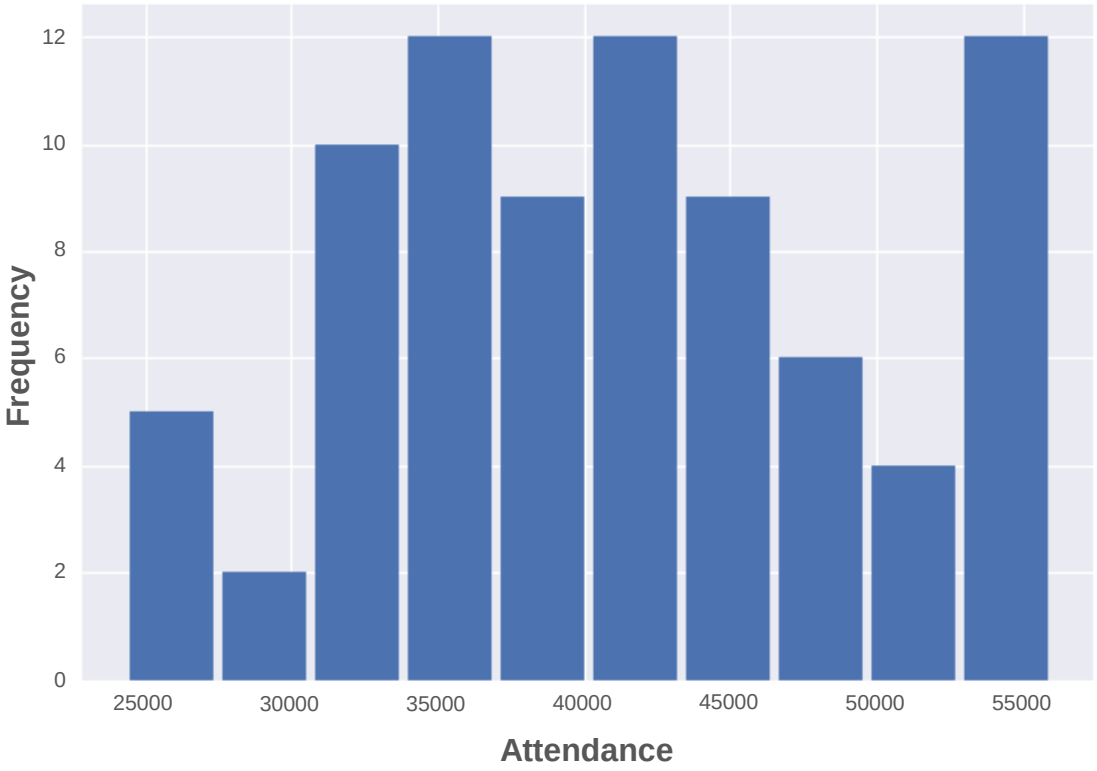
# Describing Visually



# Describing Visually

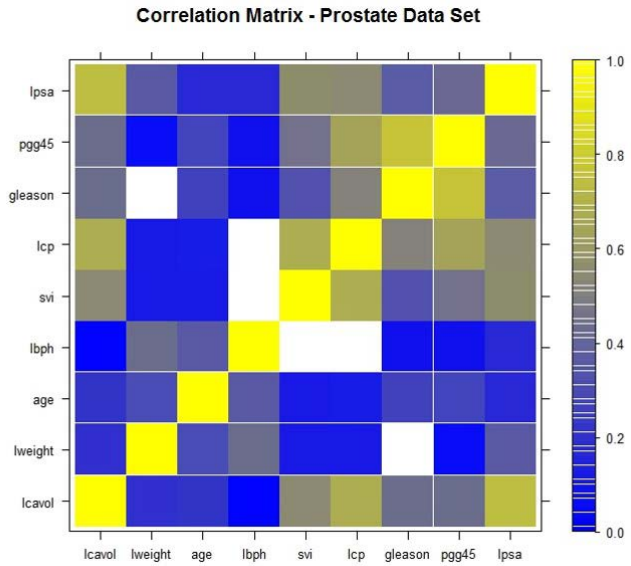


# Histogram



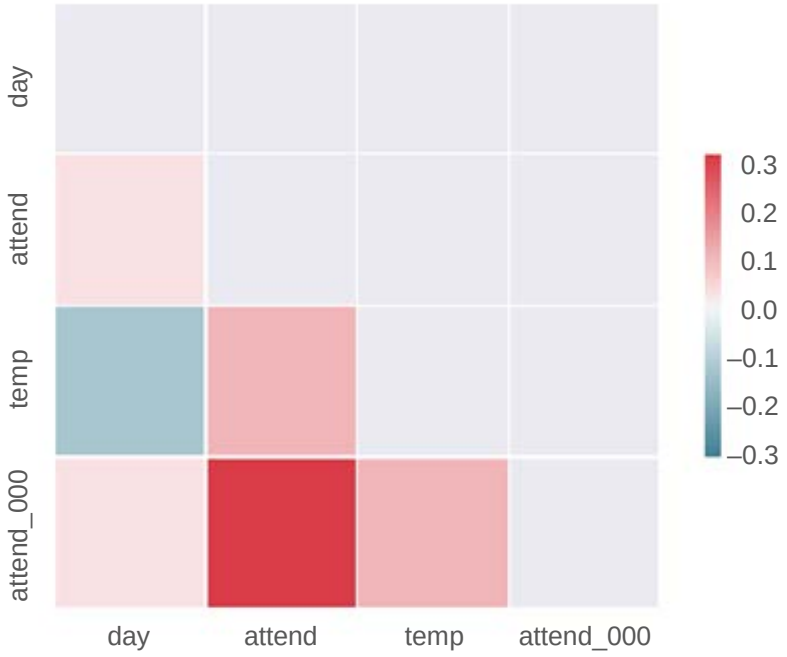

```
plt.hist(dodgers['attend'],
         normed = False,
         stacked = False,
         rwidth = .9)
plt.title("Attendance Histogram")
plt.xlabel('Attendance')
plt.ylabel('Frequency')
plt.show()
```

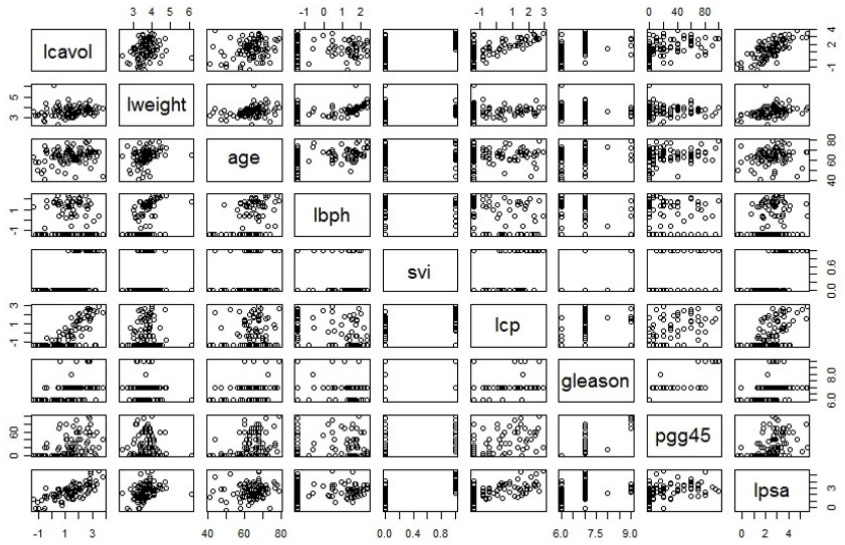
# Correlation




| | lcavol | lweight | age | lbph | svi | lcp | gleason | pgg45 | lpsa |
|---|---|---|---|---|---|---|---|---|---|
| lcavol | 1.000 | 0.194 | 0.225 | 0.027 | 0.539 | 0.675 | 0.432 | 0.434 | 0.734 |
| lweight | | 1.000 | 0.308 | 0.435 | 0.109 | 0.100 | -0.001 | 0.051 | 0.354 |
| age | | | 1.000 | 0.350 | 0.118 | 0.128 | 0.269 | 0.276 | 0.170 |
| lbph | | | | 1.000 | -0.086 | -0.007 | 0.078 | 0.078 | 0.180 |
| svi | | | | | 1.000 | 0.673 | 0.320 | 0.458 | 0.566 |
| lcp | | | | | | 1.000 | 0.515 | 0.632 | 0.549 |
| gleason | | | | | | | 1.000 | 0.752 | 0.369 |
| pgg45 | | | | | | | | 1.000 | 0.422 |
| lpsa | | | | | | | | | 1.000 |

# Correlation






School of Information Studies
Syracuse University

# Modeling

School of Information Studies
Syracuse University

# So What Is a Model?

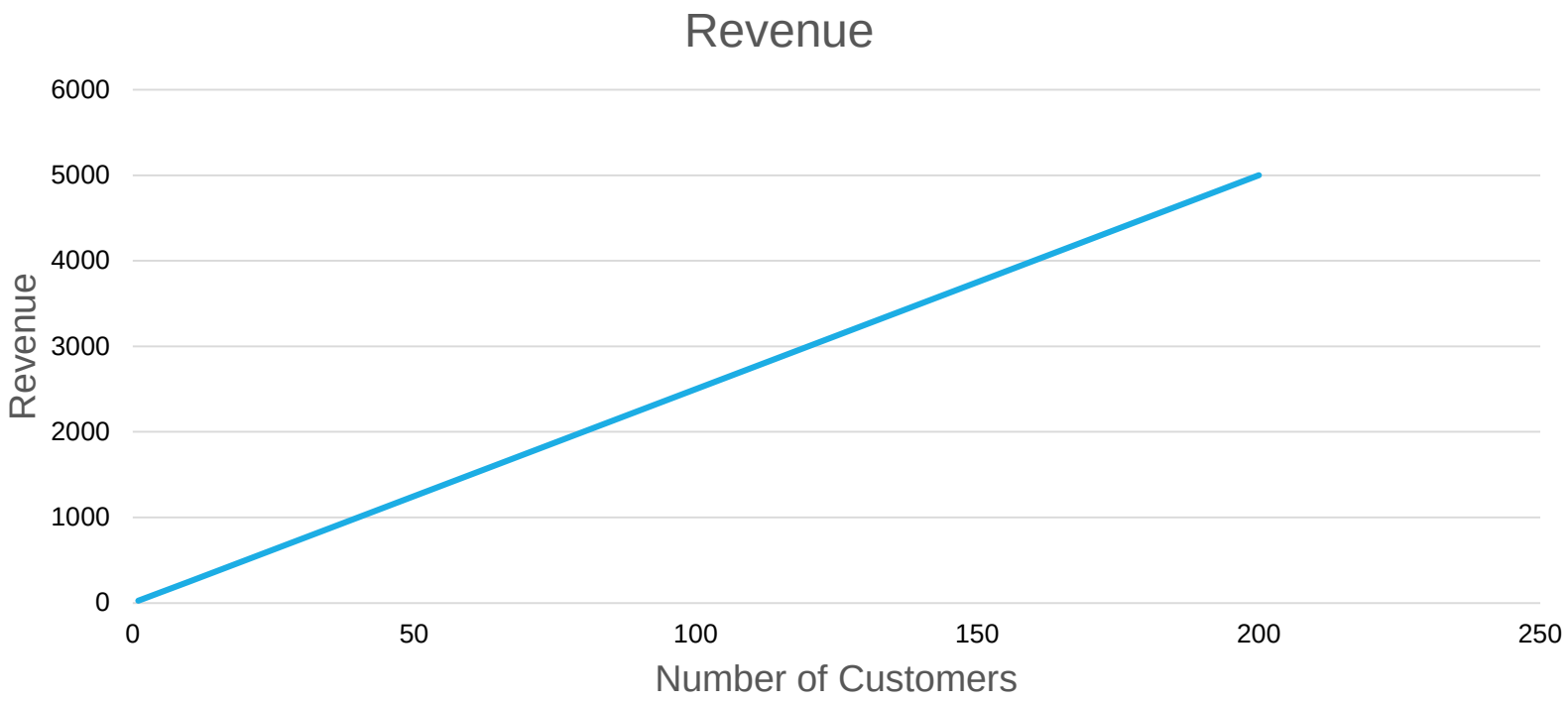
- Linear models are central to the practice of analytics
- The foundation of a broad range of analytical techniques

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

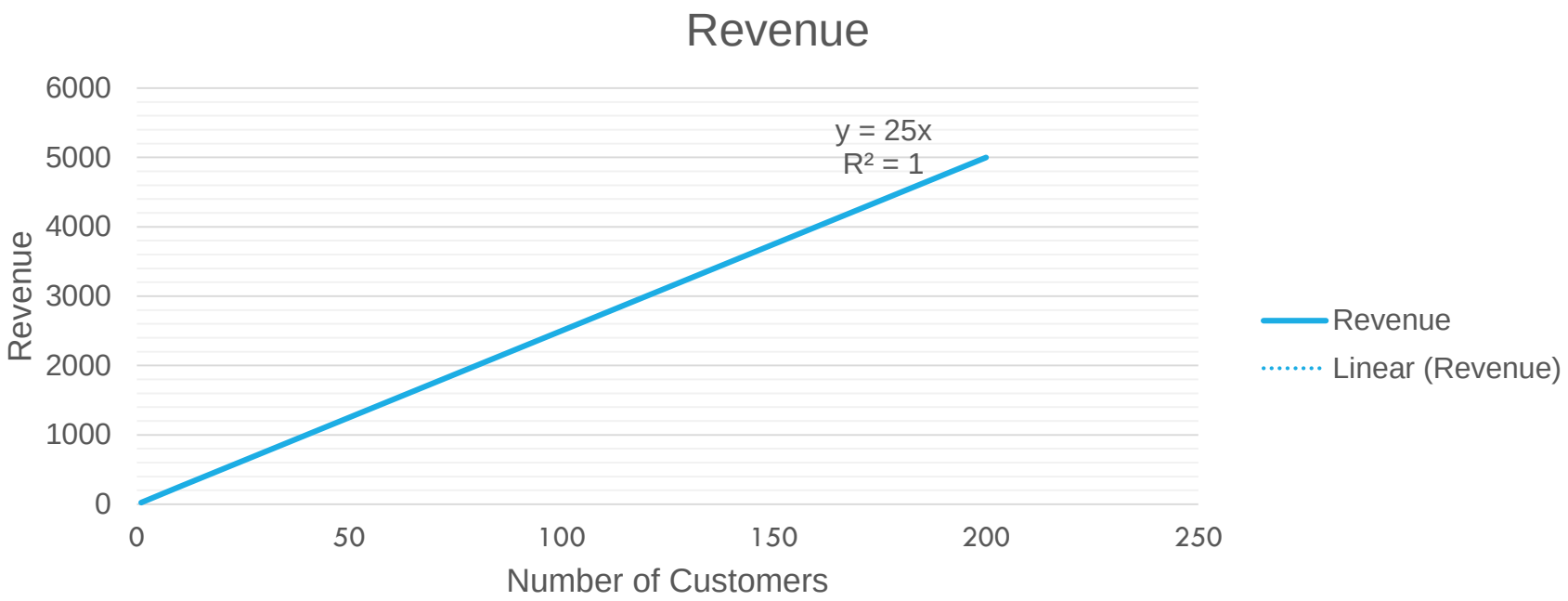
# So What Is a Model? (cont.)

| Number of Customers | Revenue |
|---|---|
| 1 | 25 |
| 10 | 250 |
| 100 | 2500 |
| 200 | 5000 |

# So What Is a Model? (cont.)



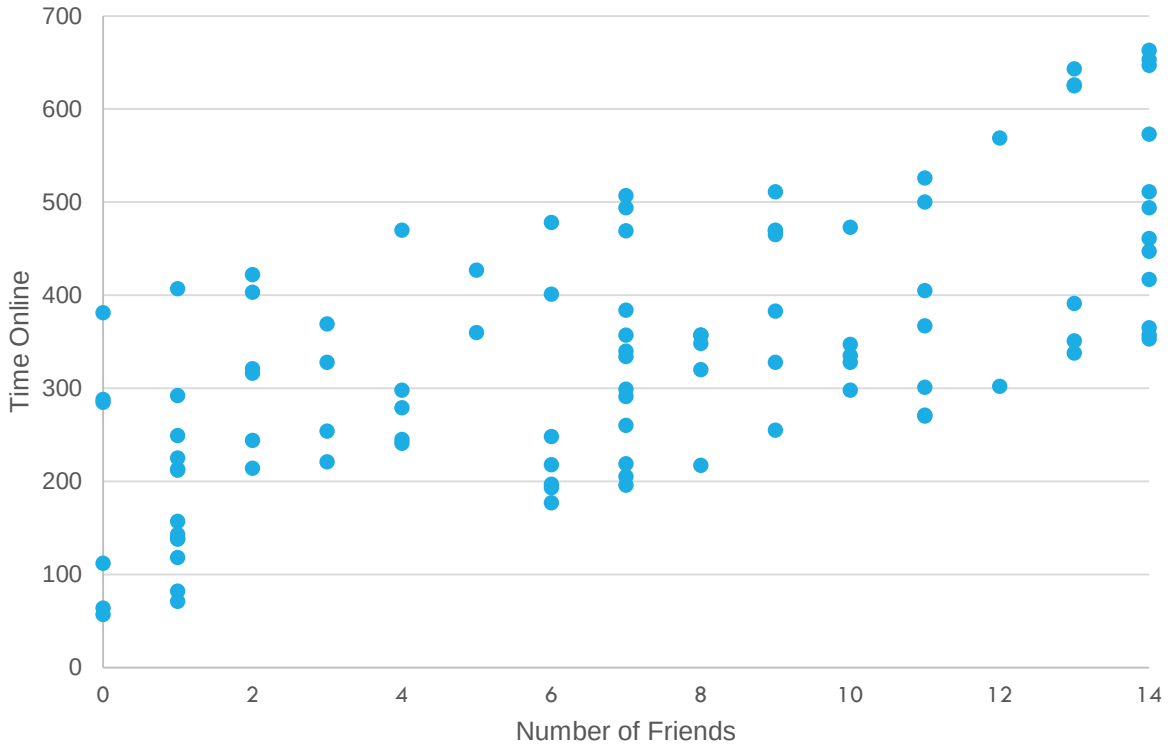
# So What Is a Model – Simple?



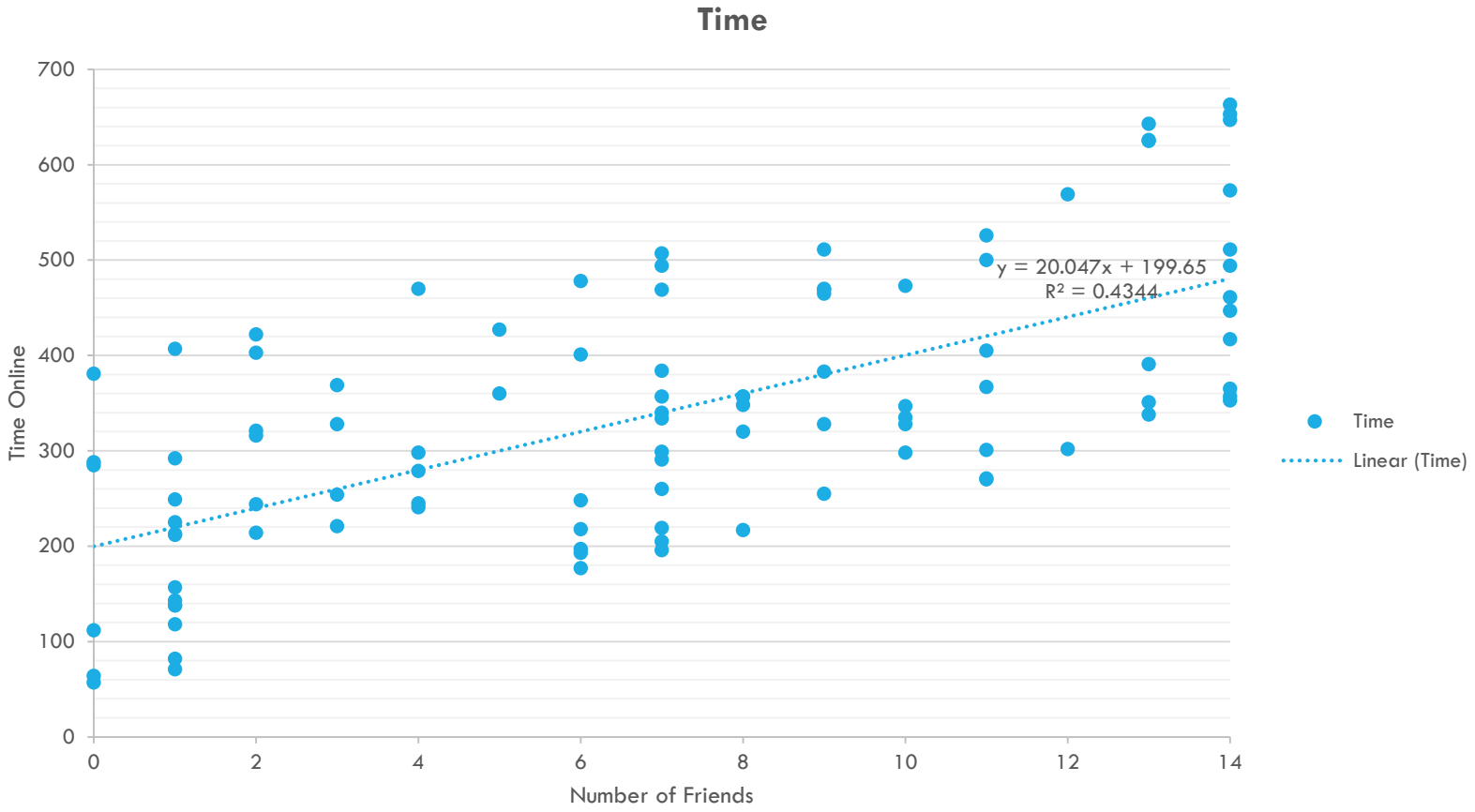
# So What Is a Model?

| Friends | Time Online |
|---|---|
| 8 | 375 |
| 14 | 657 |
| 7 | 484 |
| 7 | 381 |
| 10 | 589 |
| 8 | 394 |
| 12 | 476 |
| 3 | 319 |
| 8 | 433 |
| 9 | 529 |
| 5 | 311 |

# So What Is a Model – More Likely?



# So What Is a Model – More Likely? (cont.)






School of Information Studies
Syracuse University

# Modeling

School of Information Studies
Syracuse University

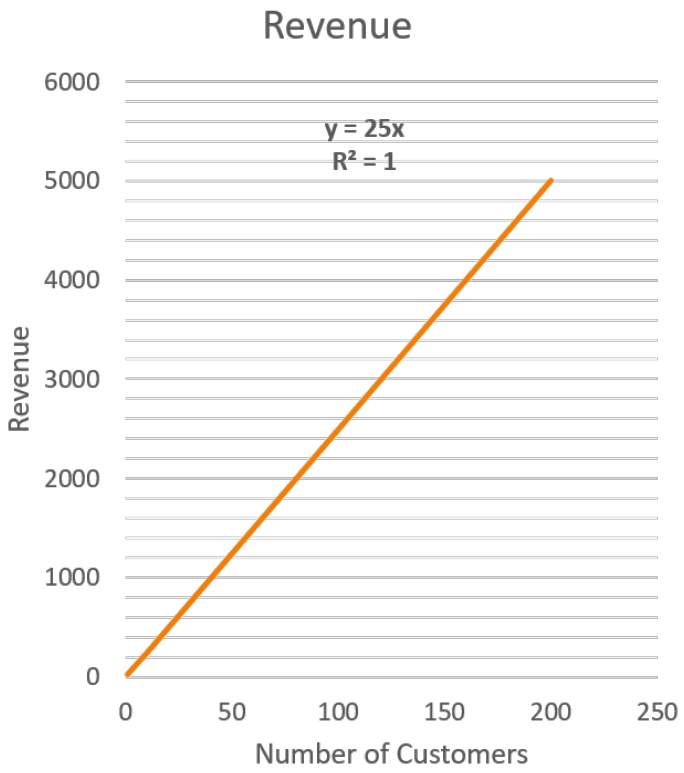
$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}$$

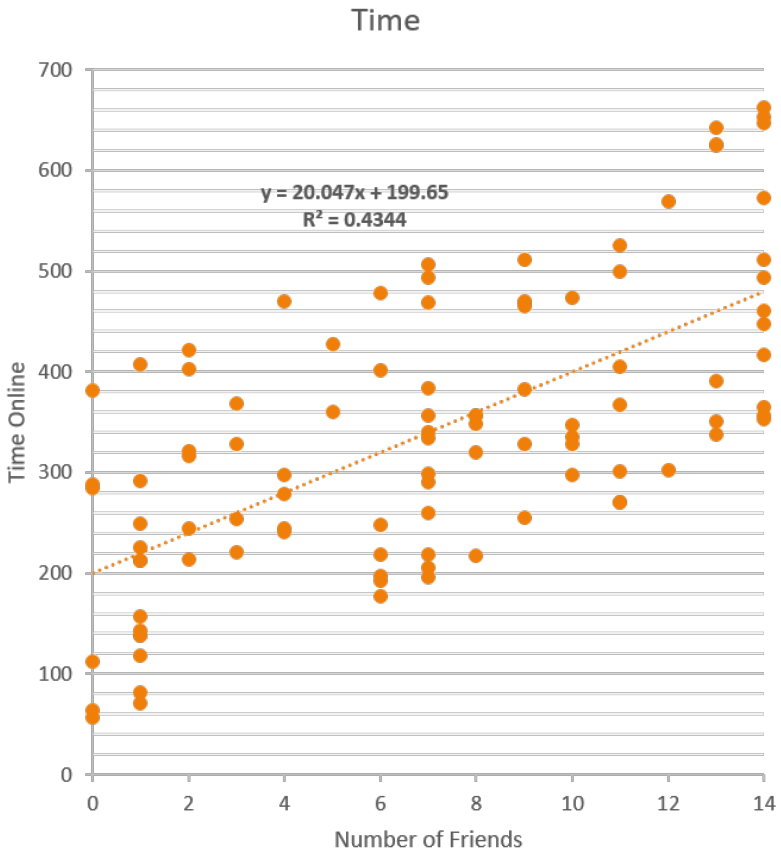
OLS Regression Results

==========

| | | | |
|---|---|---|---|
| Dep. Variable: | attend | R-squared: | 0.639 |
| Model: | OLS | Adj. R-squared: | 0.530 |
| Method: | Least Squares | F-statistic: | 5.864 |
| Date: | | Prob (F-statistic): | 4.70e-06 |
| Time: | | Log-Likelihood: | -566.87 |
| No. Observations: | 57 | AIC: | 1162. |
| Df Residuals: | 43 | BIC: | 1190. |
| Df Model: | 13 | | |
| Covariance Type: | nonrobust | | |

# R-Squared






School of Information Studies
Syracuse University

# P-Values

- Low p-value?
  - Highly unlikely to occur randomly, therefore significant
- High p-value?
  - Coefficient might actually be zero, therefore consider removing from model

```
lm(formula = hardness ~ dens, data = hardness)

Residuals:
Min      1Q  Median      3Q     Max
-338.40  -96.98  -15.71   92.71  625.06

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -1160.500    108.580  -10.69 2.07e-12 ***
dens           57.507      2.279   25.24  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 183.1 on 34 degrees of freedom
Multiple R-squared:  0.9493,
Adjusted R-squared:  0.9478
F-statistic:   637 on 1 and 34 DF,  p-value: < 2.2e-16
```

# Model Validation

| Collect | Compare | Split |
| --- | --- | --- |
| Collect new data | Compare the results with:<br>• Theoretical expectation (how much should a 0-bedroom house cost?)<br>• Earlier empirical studies<br>• Simulation (see Chapter 9 examples from text) | Split the original data with one portion for training and one for testing |