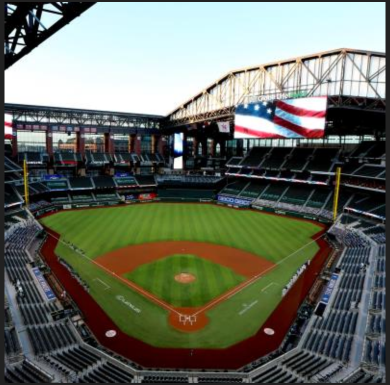
How External Factors Influence the Flow of Baseball Games

Josh De Leeuw, Michael Raybetz, and Dezmon Nash

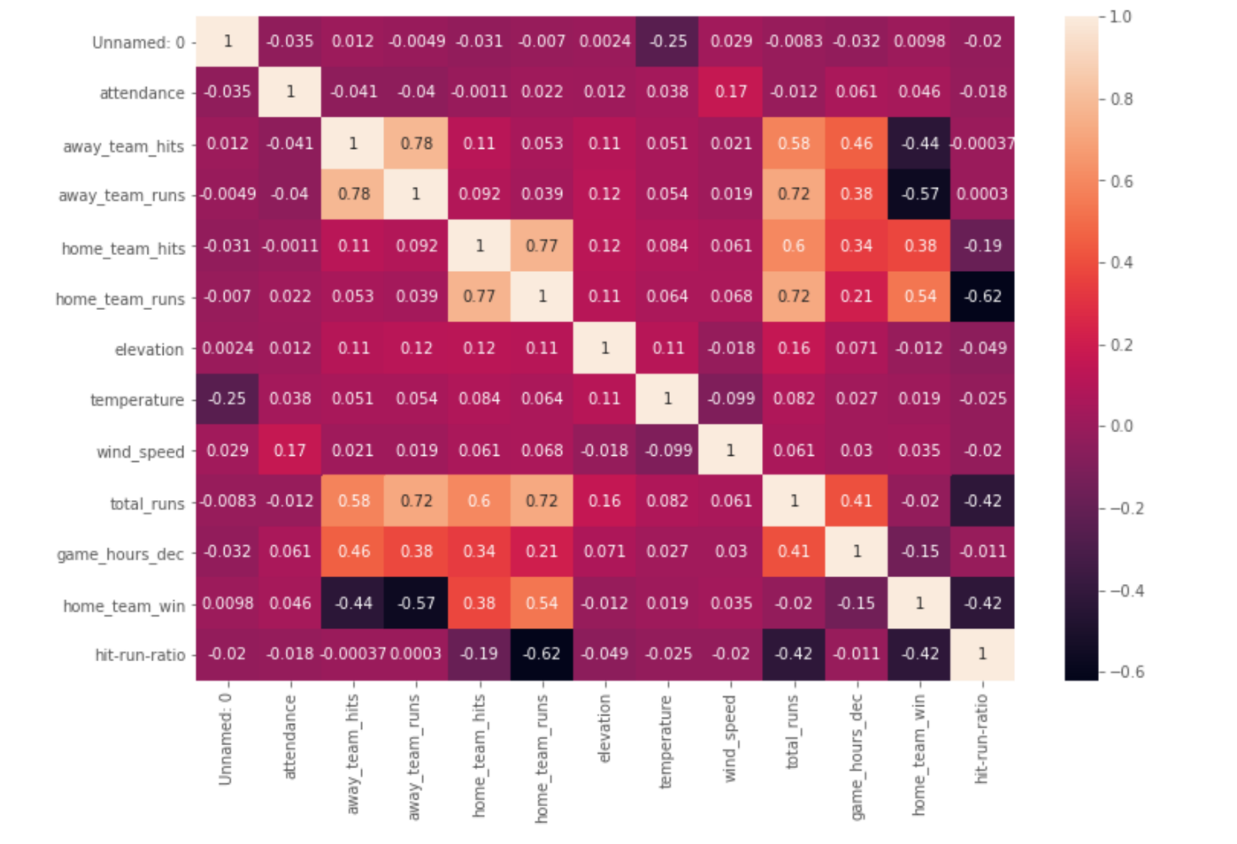


Tools and Technologies

- We received our data from Kaggle (and Baseball Reference), and used various Python libraries (Pandas, Numpy, Matplotlib) to produce the visualizations shown in the following slides

- The data was organized and the graphs were made in a Jupyter Notebook

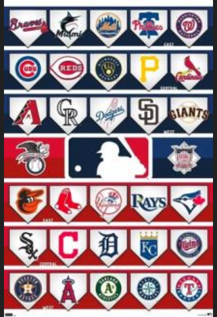
Correlation Matrix



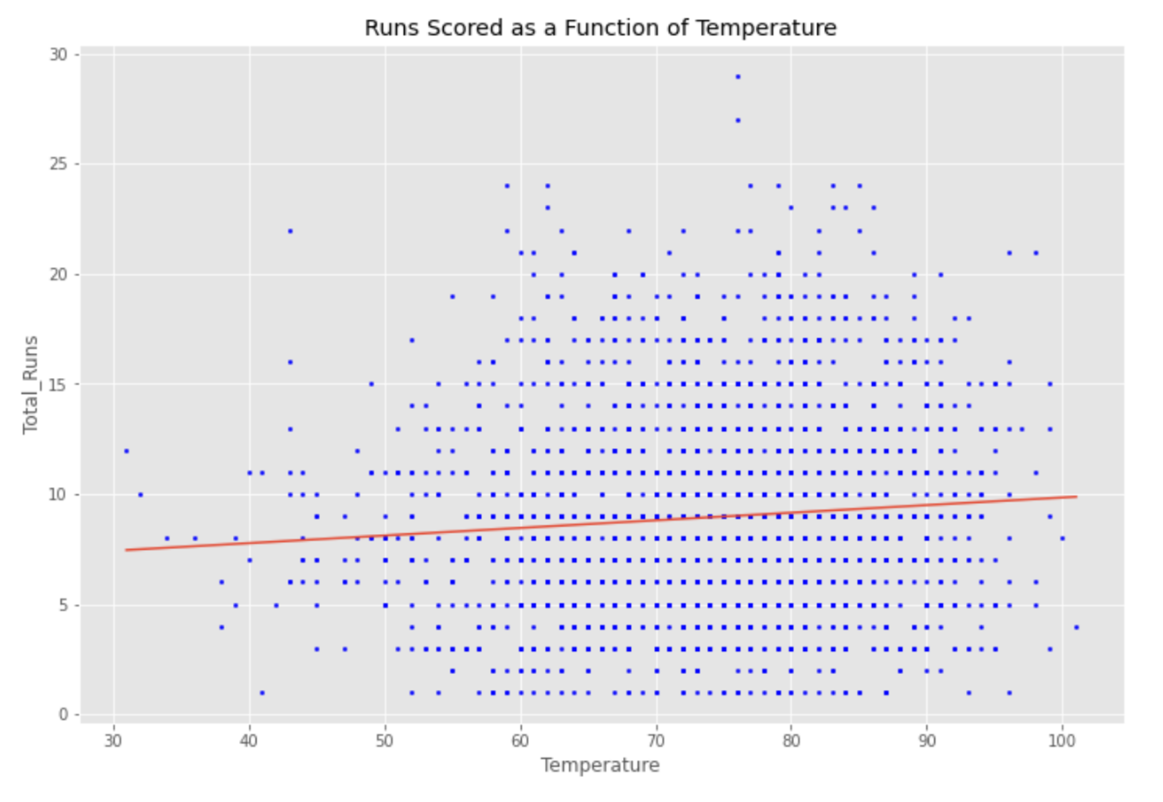
	Unnamed: 0	attendance	away_team_hits	away_team_runs	home_team_hits	home_team_runs	elevation	temperature	wind_speed	total_runs	game_hours_dec	home_team_win	hit-run-ratio
Unnamed: 0	1	-0.035	0.012	-0.0049	-0.031	-0.007	0.0024	-0.25	0.029	-0.0083	-0.032	0.0098	-0.02
attendance	-0.035	1	-0.041	-0.04	-0.0011	0.022	0.012	0.038	0.17	-0.012	0.061	0.046	-0.018
away_team_hits	0.012	-0.041	1	0.78	0.11	0.053	0.11	0.051	0.021	0.58	0.46	-0.44	-0.00037
away_team_runs	-0.0049	-0.04	0.78	1	0.092	0.039	0.12	0.054	0.019	0.72	0.38	-0.57	0.0003
home_team_hits	-0.031	-0.0011	0.11	0.092	1	0.77	0.12	0.084	0.061	0.6	0.34	0.38	-0.19
home_team_runs	-0.007	0.022	0.053	0.039	0.77	1	0.11	0.064	0.068	0.72	0.21	0.54	-0.62
elevation	0.0024	0.012	0.11	0.12	0.12	0.11	1	0.11	-0.018	0.16	0.071	-0.012	-0.049
temperature	-0.25	0.038	0.051	0.054	0.084	0.064	0.11	1	-0.099	0.082	0.027	0.019	-0.025
wind_speed	0.029	0.17	0.021	0.019	0.061	0.068	-0.018	-0.099	1	0.061	0.03	0.035	-0.02
total_runs	-0.0083	-0.012	0.58	0.72	0.6	0.72	0.16	0.082	0.061	1	0.41	-0.02	-0.42
game_hours_dec	-0.032	0.061	0.46	0.38	0.34	0.21	0.071	0.027	0.03	0.41	1	-0.15	-0.011
home_team_win	0.0098	0.046	-0.44	-0.57	0.38	0.54	-0.012	0.019	0.035	-0.02	-0.15	1	-0.42
hit-run-ratio	-0.02	-0.018	-0.00037	0.0003	-0.19	-0.62	-0.049	-0.025	-0.02	-0.42	-0.011	-0.42	1

Interesting Takes from the Matrix

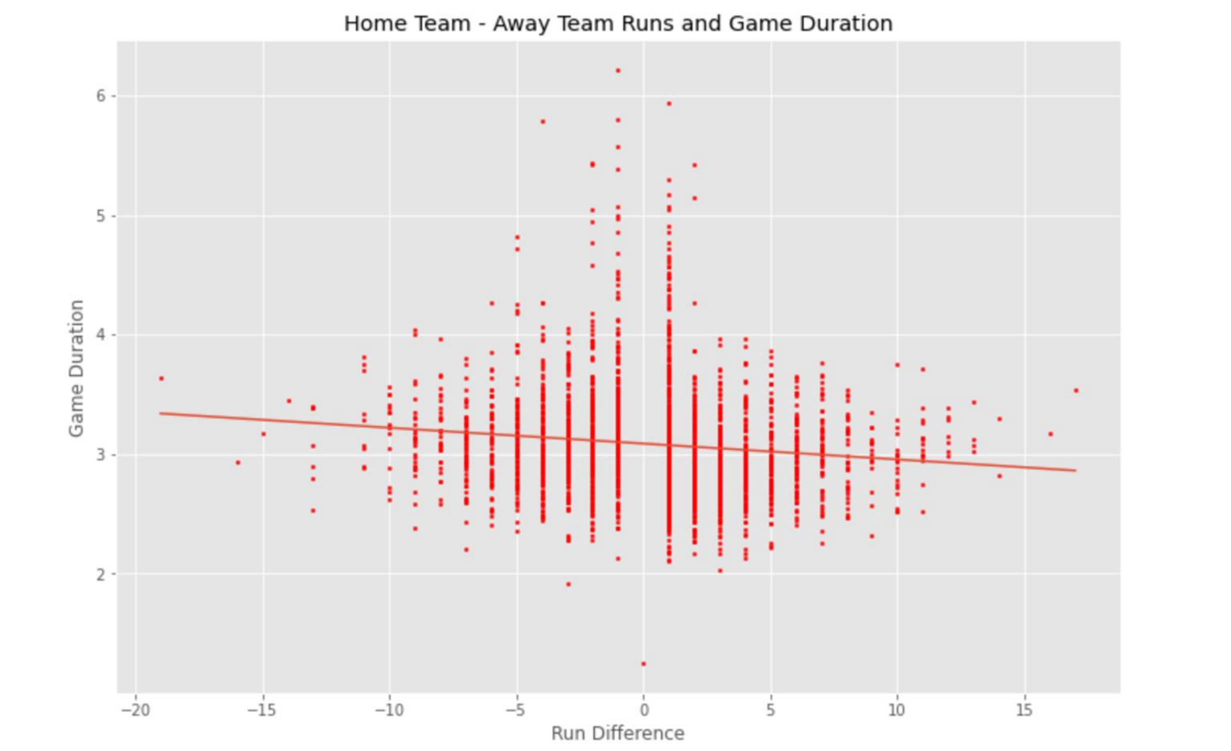
- Most factors regarding external factors (e.g. temperature, wind speed, crowd size, etc…) were not particularly correlated
  - This does not include things like “home team runs” vs “home team hits”, which of course have strong dependencies
- Nevertheless, there were some interesting trends (and non-trends), that we will take you through in the following slides



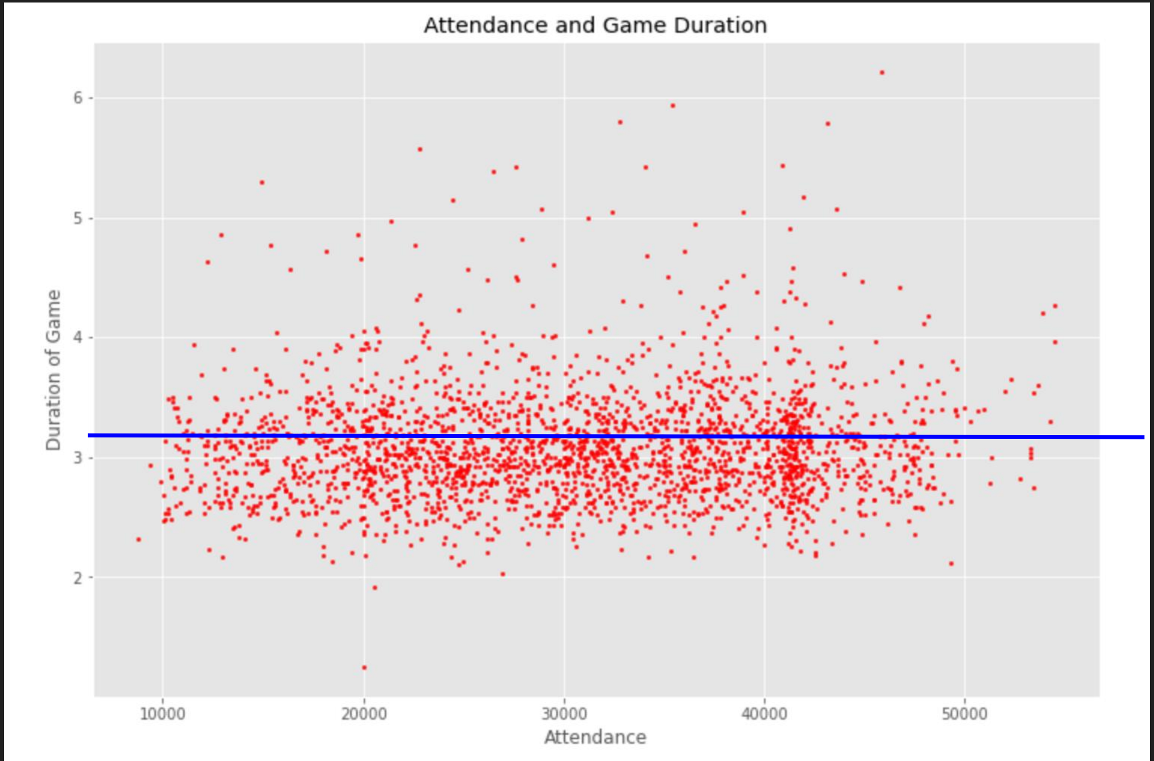
Runs Scored vs. Temperature



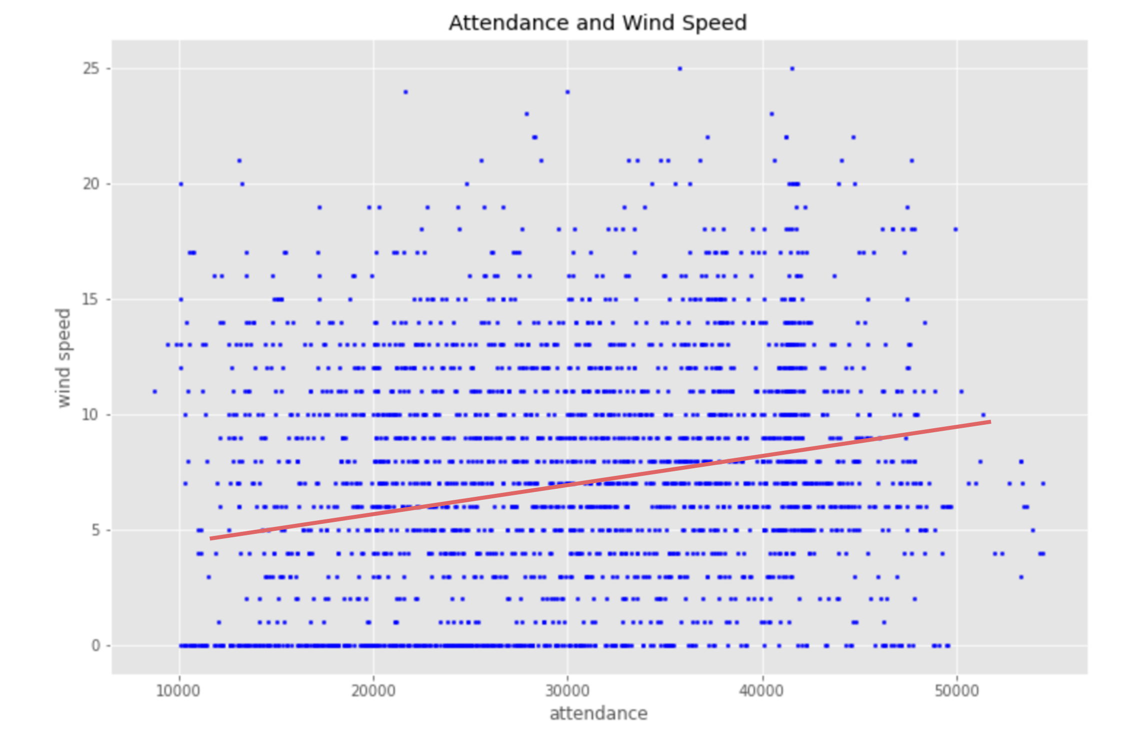
How Run Difference Affects the Length of Game



Do more attended games last longer?



Surprising Correlation: Attendance vs Wind Speed



Conclusions From Graphs

- Even when correlations were present, it can be hard to prove causation
    - Either factor could be causing the other, or a separate factor may be causing both of them
    - My favorite statistical example: ice cream consumed per month and drownings per month are correlated, even though eating ice cream clearly does not cause one to drown
        - Hot weather causes both of them!
- However, some correlations we found were runs scored vs. temperature, attendance vs. wind speed, and run difference vs. the length of game
- Further analysis could involve working with different Kaggle datasets, or trying to dig further into this one