[MLB] Game Results 2012-2021

TEAM 7 - Adam Chow, Jaelyn Do, Chris Kim, Jessica Li, Kevin Zhang

Introduction & Dataset Summary:

Introduction:

- Interest: finding effects of different variables on winning/team performance
- Dataset from Kaggle: MLB Team Ranks & Game Results from 2012-2021 | Kaggle
  - Data scraped off baseball-reference.com
  - 2 separate csv files (mlb_games.csv and team_ranks.csv), we chose to look at mlb_games.csv only
  - 88 columns total (25 from mlb_games.csv and 63 from team_ranks.csv), we used 7 from mlb_games.csv

Research Questions:

- Does the home-field advantage actually exist?
- Does geography affect how many runs a team scores?
- Does time of year affect how many runs are scored on average?

Data Preparation and Cleaning:

Combine (grepl) - win_or_lose
- W-wo, L-wo, W &X, L &X, etc.
  - wo: walk-off
  - W-wo, W &X —> W
  - L-wo, L &X... —> L

Region
- Mapping team to corresponding home state

Total Runs
- Combined Runs and Runs Against

Date.dec
- Date -> Date.dec

Avg Runs per month
- Mean runs per listed calendar month

Data Summaries

Does geography affect team performance?- Region

arizona	california	colorado
695	3489	702
district of columbia	florida	georgia
700	1389	698
illinois	maryland	massachusetts
1401	698	703
michigan	minnesota	missouri
694	698	1393
new york	ohio	pennsylvania
1391	1387	1395
texas	washington	wisconsin
1397	700	701

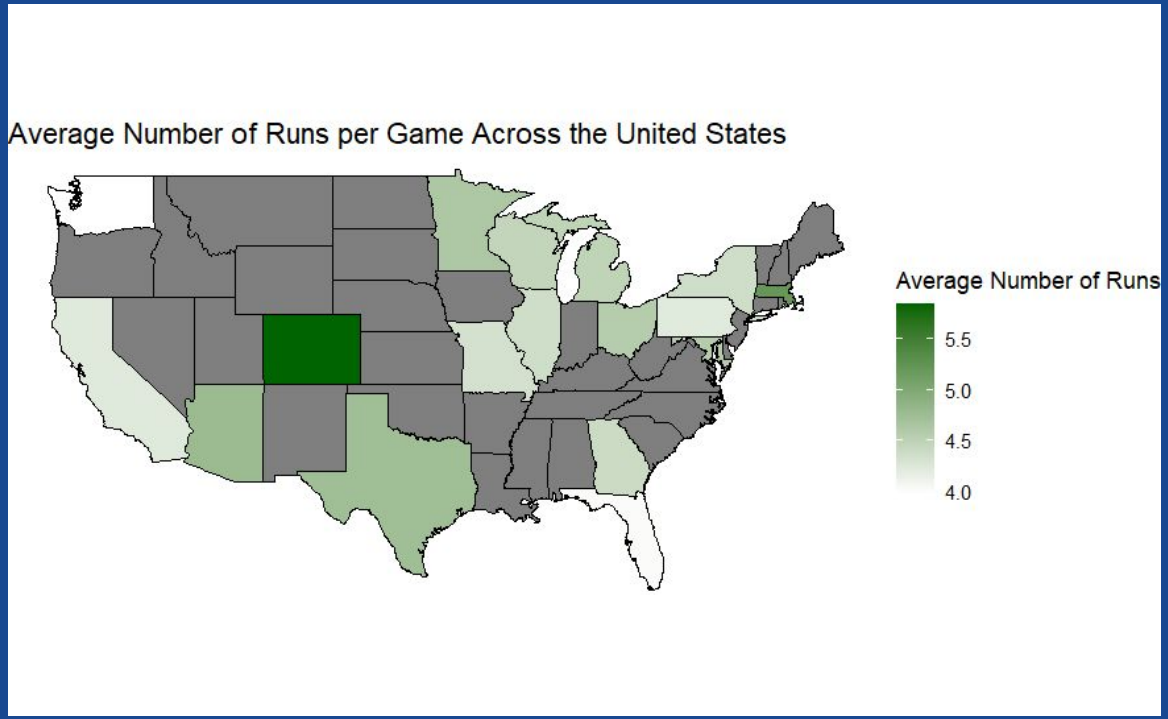
Does the home-field advantage actually exist? - Win_or_Lose

	away	home
L	11184	9734
W	9734	11184

Does time of year affect how many runs are scored on average?- Average Runs per month

	month	mean_runs	min_runs	q1_runs	median_runs	q3_runs	max_runs
1	3	4.53	0.00	2.50	3.92	6.00	14.00
2	4	4.33	2.64	3.85	4.27	4.78	6.80
3	5	4.40	2.82	3.89	4.31	4.89	7.04
4	6	4.45	2.22	3.85	4.44	5.00	6.77
5	7	4.44	2.50	3.88	4.40	4.95	7.25
6	8	4.49	2.78	4.00	4.41	4.89	7.14
7	9	4.41	2.38	3.84	4.37	4.96	6.88
8	10	3.69	0.00	2.00	3.00	5.00	14.00

Question: Does geography affect team performance?



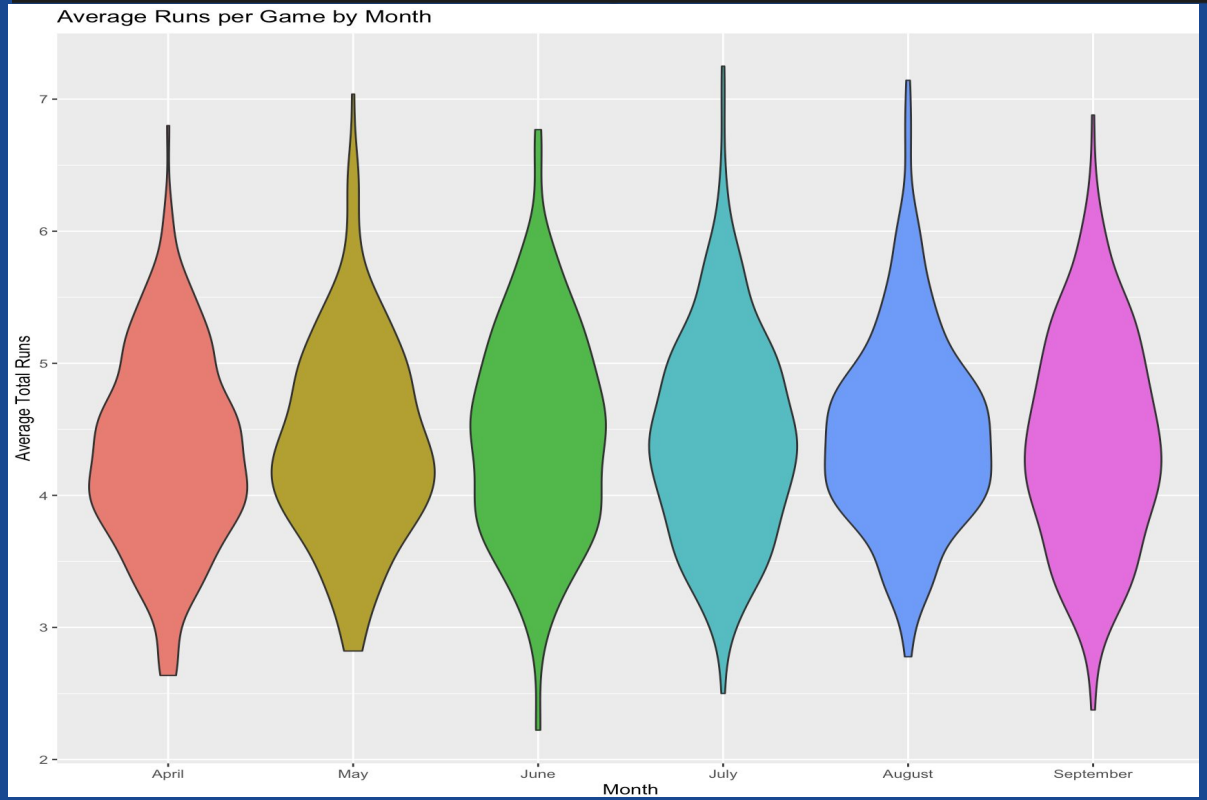
Observations:

- Less runs near the coasts
- More runs in the South
- Massachusetts outlier

Conclusion:

- Elevation and temperature effects

Question: Does time of year affect total run production?



Violin Plot:

- Discrete X variable (Month)
- Continuous Y Variable (Average Total Runs)

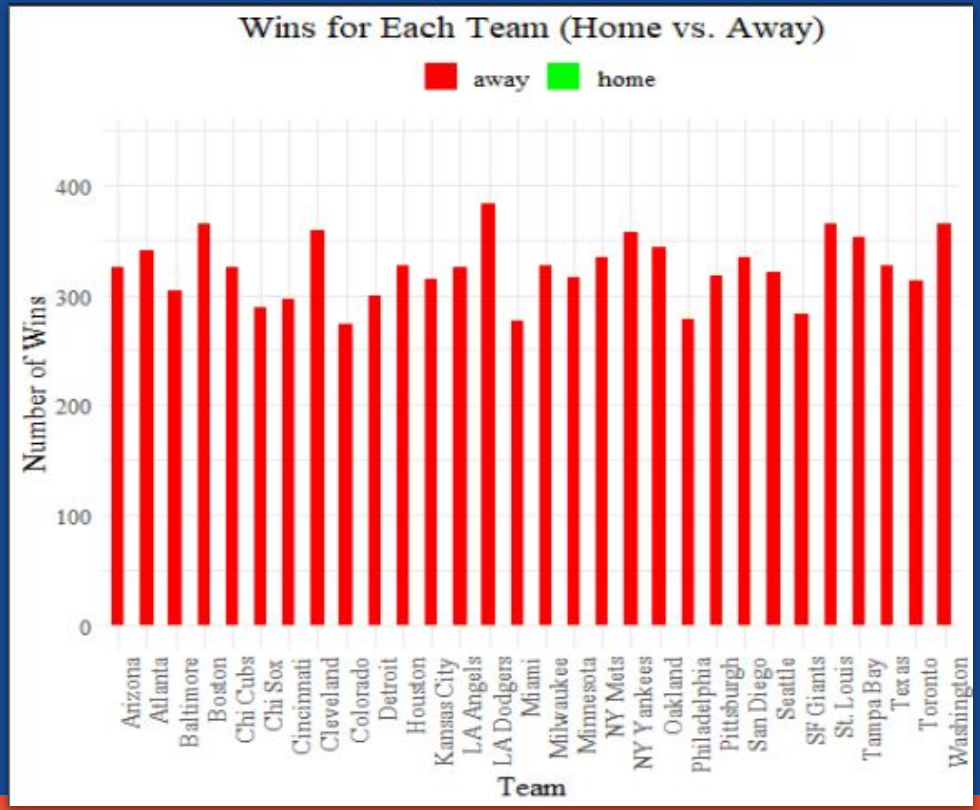
Observations:

- Higher peaks in June, July
- Lower bodies for April, September
  - Temperatures are lower
  - Higher chance for inclement weather

Conclusions:

- Month as a single factor has little impact on run production
- Need other factors to attribute correlation

Question: Does home-field advantage exist?



	away	home
L	11184	9734
W	9734	11184

Bar Graph: comparing numerical data between different groups and changes over time

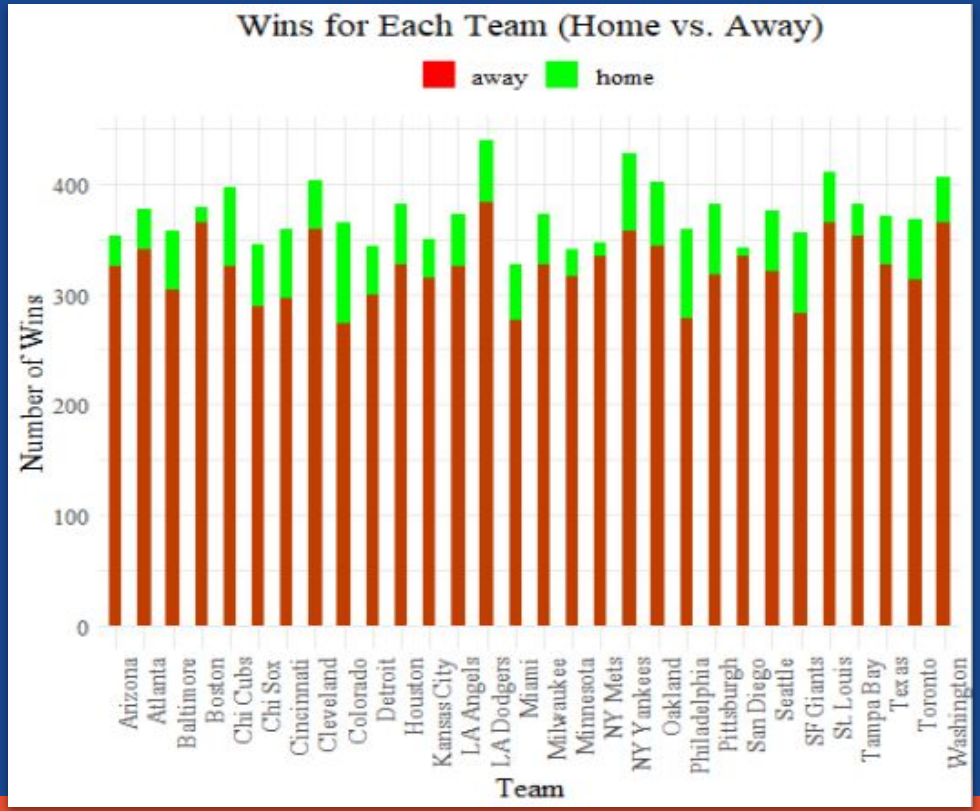
Observations:

- More wins across every team during home games
- Total difference = 1,450 more wins from home games than away games

Conclusion: The home-field advantage does exist in the scope of this dataset.

Graph of the total number of wins for away and home games for each team between 2012-2021 seasons

Question: Does home-field advantage exist?



	away	home
L	11184	9734
W	9734	11184

Bar Graph: comparing numerical data between different groups and changes over time

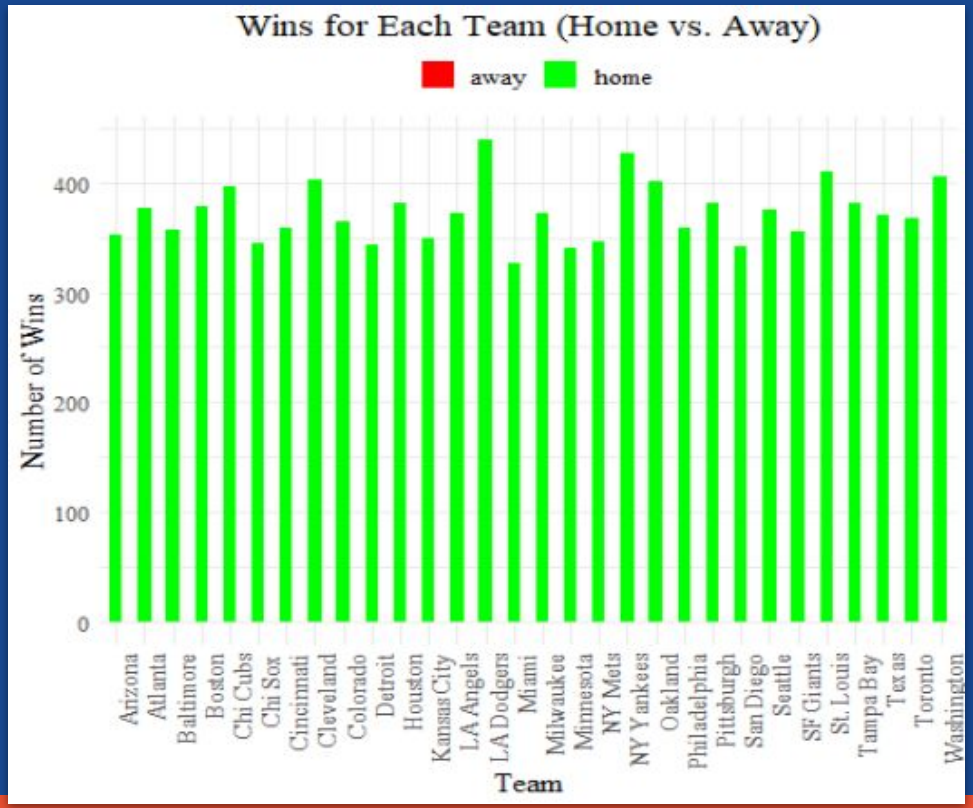
Observations:

- More wins across every team during home games
- Total difference = 1,450 more wins from home games than away games

Conclusion: The home-field advantage does exist in the scope of this dataset.

Graph of the total number of wins for away and home games for each team between 2012-2021 seasons

Question: Does home-field advantage exist?



	away	home
L	11184	9734
W	9734	11184

Bar Graph: comparing numerical data between different groups and changes over time

Observations:

- More wins across every team during home games
- Total difference = 1,450 more wins from home games than away games

Conclusion: The home-field advantage does exist in the scope of this dataset.

Graph of the total number of wins for away and home games for each team between 2012-2021 seasons