Ashley Marietta, Gianna Gordulic, Molly Kenefick
Professor Barron
MSBR-70260-03
9 October 2025

The "Caitlin Clark Effect" on the WNBA

Introduction and Motivation

Dubbed the "Caitlin Clark Effect" by Shelley Lucas from the Department of Kinesiology at Boise State University, Caitlin Clark's collegiate success raised interest in women's college basketball as she led the University of Iowa's basketball team to long NCAA championship tournament runs, making it to the national final in both 2023 and 2024 (Judge and Petersen; "Iowa"). Though data has been collected for her college years, there is yet to be evidence or a formal study of Clark's influence on WNBA attendance. This poses the interesting question of whether one player can drive fan engagement with a sport, particularly with Clark bringing national attention to women's basketball. If so, Clark's influence could lead to exposure for other female professional sports and athletes, driving ticket sales, media deals, and product endorsements. In addition, understanding the impact of Caitlin Clark could inform marketing efforts, team strategy, and WNBA economics. This question is especially relevant in 2025, as the WNBA has seen record-breaking viewership and attendance (Associated Press). Similar to quantifying the "Caitlin Clark Effect" during her college years, this analysis aims to predict the impact of Caitlin Clark's presence on WNBA attendance.

Related Work

Looking at the average game attendance through her college seasons, a study by Lawrence W. Judge of Ball State University and Jeffrey C. Peterson of Baylor University shows that fan attendance rose from 5,387 before Clark's freshman year to 13,877 by her final season. Even after Clark left for the WNBA, the average remained at 9,890, demonstrating a sustained interest (Judge and Petersen). Their study utilized cross-sectional data spanning six NCAA seasons for women's basketball, tracking attendance using game-level data from box scores for one year prior to Caitlin Clark's arrival at the University of Iowa, four years during her career, and one year after her graduation. Key variables included attendance, percent capacity, location, and Caitlin Clark's presence on the court. Researchers used one-way ANOVA tests to find that the mean attendance during each year of Clark's career increased and remained higher than in 2019, even after Clark's graduation. Judge and Petersen state, "Overall, these findings demonstrate clear, statistically robust growth in both attendance and venue capacity utilization during and following Caitlin Clark's career. The trajectory of fan engagement suggests that the Caitlin Clark Effect produced not only a short-term surge in interest but also a structurally significant and potentially enduring shift in consumer behavior surrounding women's collegiate basketball" (31-34).

Other less extensive studies have also been conducted, demonstrating Clark's off-the-court impact on the economy. For example, the average price for Iowa women's basketball games increased by 224 percent during her senior year, and she raised the Iowa state GDP by over $52 million, according to the Common Sense Institute of Iowa during her four years as a Hawkeye (McGuire; Zuritsky).

Data Description

For our project, we have decided to work with a WNBA dataset from the 2022 through 2025 seasons from the Wehoop library. This dataset consists of 57 variables. We originally used this to get Caitlin Clark's minutes played, points scored, three-point field goals made, date, game type, and day of the week the game was played. We are also utilizing a dataset from *Data Explorer - Across the Timeline - Stats, Facts, and Memories from the Storied History of Women's Basketball* to see the attendance from 2022 to 2025. We used the variables of city, attendance, and home/away teams. We first added a binary variable to see if Caitlin played or not. We also adjusted the date format of the datasets to make them easier to merge. Then, we filtered the WNBA dataset to show Caitlin Clark's statistics when she played, merging our attendance dataset and the new Caitlin Clark dataset to show each game, the date of the game, the day of the week, the game played, the game type, the home team, the away team, the city, and the attendance. As mentioned, we used the WeHoop library to get Caitlin Clark's minutes, points, and three-point shots made. We used these variables to create some of our variables in our dataset. These included the variables "Caitlin_points_lag3," "Caitlin_minutes lag_3," and "Caitlin_3pts made_lag." We also added the variables "stadium_norm," "stadium_mean_prior," "stadium_rolling_k_mean_prior," "stadium_prev_game_attendance," "stadium_prior_games_count," and "stadium_mean_prior_filled." This came out to 18 columns and 1,046 rows. We created two visualizations after cleaning and collecting the data. Figure 13 is a line graph that shows attendance when Clark played versus when she did not, demonstrating that Clark had a significant impact on attendance. Figure 10 is a line graph that shows attendance when Clark does or does not play during different types of games. Regular-season games present the highest average attendance when she does play, which can prove her impact once again.

We faced a few key challenges along the way, including inconsistent date formatting across the datasets that required manual cleaning before automating. Another challenge was aligning Caitlin Clark's game data with attendance figures, since they came from separate sources. At first glance, an important statistic to note is that the maximum attendance from 2022 to 2025 was 20,711 (found using the MAX function in Excel), when Caitlin and the Fever played the Washington Mystics in D.C. In gathering data insights like this during collection, we were able to mitigate these problems throughout the project.

Methods

After cleaning the data and using code to add lagged and other variables (capturing past attendance, "Caitlin_points_lag3," "Caitlin_minutes_lag3," "Caitlin_3pts_made_lag3," "stadium_norm,

stadium_mean_prior," "stadium_rolling_k_mean_prior," "stadium_prev_game_attendance," and "stadium_prior_games_count"), our team began running several statistical tests and models to determine which one was the best at predicting attendance – specifically considering Caitlin Clark's influence, if any. These models included XGBoost, Bagging, and Linear Regression. Running these three models enabled us to compare linear and nonlinear approaches, evaluate their accuracy, and determine which method best captured the complex relationships in our dataset.

Linear Regression is often more simplistic than other methods and is great for interpretability and establishing a clear baseline for comparison. However, it often fails to capture nonlinear effects and complex relationships among variables, such as how attendance can be influenced by multiple factors (player performance, stadium characteristics, etc.), which is the reason we also used both XGBoost and Bagging. These methods are better suited for modeling nonlinear patterns and multiple interactions. Bagging is a technique that improves accuracy, reduces variance, and avoids overfitting by averaging results across decision trees, while XGBoost (Extreme Gradient Boosting) modeling uses sequential tree building and gradient-based optimization to minimize prediction errors and capture complex nonlinear relationships. The issue with these two techniques is that both require careful hyperparameter tuning and validation to prevent overfitting. It should be noted that, in the case of "The Caitlin Clark Effect," these models are particularly important because attendance patterns likely involve many nonlinear relationships and complex interactions among variables.

For tuning purposes, we used GridSearchCV with cross-validation to identify the optimal hyperparameters for each model. The dataset was split into training and test sets, with the models trained on the training dataset and evaluated on the test set to prevent overfitting and ensure generalizability. Before adding stadium-related features, we also trained and evaluated the models on the original dataset to establish a baseline performance. From the tuned models, we visualized several important evaluation metrics, including the R^2 value, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), comparing results from both the untuned and tuned models. These values provided evidence for selecting the best-performing model to use in further analysis.

Finally, for model explainability, the group used SHAP (SHapley Additive exPlanations) values to help determine which features had the greatest influence on the XGBoost model's predictions for game attendance. This method is instrumental in interpreting complex models, allowing us to identify the most influential factors like those related to Caitlin's impact on league attendance. Additionally, we highlighted these features through visualizations, producing meaningful results and conclusions.

Results

Before tuning the models, we found that the most appropriate model was linear regression, which achieved an R^2 value of 0.14 and a MSE of 17,980,021.25, while the bagging and XGBoost models

produced negative R^2 values and high MSEs. The negative R^2 values were concerning because they indicated that the models performed worse than a simple mean-based prediction and that they were not accurately capturing the underlying relationships in the data. Because there could have been multiple factors driving attendance, it was imperative that we run a search for the optimal hyperparameters and rerun the models once again.

The hyperparameters for the most appropriate model – the *tuned* XGBoost – included the number of estimators (50, 100, 200), learning rate (0.01, 0.1, 0.2), maximum tree depth (3, 5, 7), subsample fraction (0.6, 0.8, 1.0), and a fixed random state (42) for reproducibility. These hyperparameters were optimized using GridSearchCV, which tested parameter combinations to identify those that produced the best model performance.

After comparing each untuned and tuned model and using hyperparameters, the tuned XGBoost model produced the lowest MSE (75,955,633.5), RMSE (2,609.20), and the highest R^2 value (0.35) on the test set, confirming it predicted attendance the most accurately and captured the key factors driving variation in attendance better than the other models.

Using SHAP values to interpret feature importance, the top predictor was whether Caitlin played, highlighting her strong influence on WNBA attendance. The next most important factors were "stadium_mean_prior_filled" and "stadium_prior_games_count," both reflecting typical stadium capacity and prior attendance trends, suggesting that historical fan engagement and venue popularity also meaningfully shape attendance predictions.

There were several challenges we encountered in the project and model-fitting process. Our original goal was to forecast future attendance based on Caitlin, but we lacked the data points to create such a predictive model. Caitlin has only been in the WNBA for two seasons, and between that and injuries, there is not enough evidence to reliably create a future forecast in the allotted project timeframe. In addition, we do not know the future season schedules, limiting our ability to predict attendance for upcoming games. Therefore, we pivoted to focus on her impact on WNBA attendance and used SHAP features to see which had the greatest influence on the models.

Another challenge, which was previously mentioned, was the initial poor performance of the untuned models, which produced negative R^2 values, suggesting that they were highly inaccurate. We were able to overcome this issue for the most part when we used GridSearchCV and hyperparameters, but the models still did not achieve particularly high predictive accuracy. Overall, these results confirm that Caitlin's presence and consistent stadium trends are the strongest drivers of attendance, which will be discussed in the subsequent sections.

Discussion

As mentioned in previous sections, we ran three models: linear regression, XGBoost, and bagging. When looking at all of these models, it is clear that the XGBoost-tuned model was the best for our data. Our XGBoost model had the lowest Mean Squared Error (MSE) equal to 75,955,633.50, the lowest Root Mean Squared Error (RMSE) of 2,609.20, and the highest R^2 value of 0.35. We were surprised by these results, but even more so by the results of the other two models. For our bagging model, the MSE value was 9,607,382.71, the RMSE value of 2,797.12, and the R^2 value was 0.26. The linear regression had an MSE value of 14,706,457.69, RMSE value of 3,834.90, and an R^2 value of -0.39. Although it was evident that the XGBoost model performed the best, none of our results were as high as we had hoped. However, we created some visualizations with interesting insights that supported our claims surrounding Caitlin Clark and WNBA attendance.

The visualization that gave us the most insight was Figure 3. This figure is a beeswarm SHAP plot that shows the most influential features in predicting attendance. In this visualization, it is clear that "Caitlin_played" was the most important predictor for attendance. This proves our initial question that when Caitlin does play, there is a significant impact on attendance due to the high SHAP values in the plot. "Stadium_mean_prior__filled" was notable as well; this variable has to do with the stadium's average attendance. Higher attendance averages correspond to better SHAP values, while lower averages are associated with smaller SHAP scores.

Another visualization that we found interesting was Figure 9. It is a scatter plot of the attendance of the WNBA throughout the 2022-2025 seasons. The blue dots signify when Clark did not play, and the green dots are when she did. It is evident from this graph that the attendance is higher when she plays, and in games where she was not playing, the average attendance continued to increase. This proves that she has a significant impact on the WNBA and its attendance.

Figure 1 shows fascinating insights from our dataset as well. This SHAP dependence plot shows a positive relationship between "stadium_mean_prior_filled" and its SHAP value. As the historical average attendance at a stadium increases, the likelihood of higher predicted attendance increases. The red dots represent when Clark played, and the blue dots represent when she did not play. Most of the red dots are higher on the SHAP value axis compared to the blue dots, suggesting that Clark's playing and a stadium's historical attendance are crucial in predicting game attendance.

With all of these insights, we suggest the WNBA should change its marketing strategy. Caitlin Clark is one of the faces of the WNBA and has tremendously increased attendance. However, if they could add or market another young player, this could draw in more fans and lead to higher attendance.

Conclusion and Future Work

In conclusion, our report showed that Clark's presence is statistically significant in the WNBA, and the XGBoost updated model had the most accuracy when compared to the linear regression and

bagging updated models. When analyzing the conclusions from the model, it becomes clear that running multiple tests on the data is crucial to identifying the model with the highest accuracy. For example, in some of our in-class examples, bagging or linear regression performed better than other model types. Our model was used to determine Caitlin Clark's impact on WNBA attendance, and we can conclude that her presence has a significant positive effect on game turnout.

With more time and resources, the next step would be to improve the overall accuracy of the model by adding in more features, such as further past attendance data, and by tracking more of Clark's games in future seasons when she is not injured. Additionally, we would aim to build a forecasting model to predict future WNBA attendance, assuming Caitlin Clark plays in all games, though this may not hold if she misses games due to injury or other reasons. Finally, we would share the data with WNBA teams and companies looking to invest in women's athletics to support the idea that women's sports are a growing industry and a business opportunity.

Contribution

Each of the team members played an instrumental role in contributing to the project, presentation, and final paper.

Molly Kenefick helped clean the data, specifically fixing the date formatting, created the "Project 1 Code," and added lagged and other variables to the original dataset to strengthen the models. She ran each model, used GridSearchCV to identify the hyperparameters, and tuned each to determine the best-performing method. Molly also created most of the final visualizations, wrote the Methods, Results, and Contribution sections, and added the Deliverables, Results, and Methods to the slide deck. Additionally, she helped Gianna with the appendix section.

Gianna Gordulic cleaned and collected the data, including manually inputting points scored, and helped create the early visualizations. She wrote the Data Description and Discussion sections, collaborated with Molly to complete the appendix section, assisted Ashley with the bibliography, and added the Data Description to the slides.

Ashley Marietta found related studies and articles, helping her to write the Introduction and Conclusion sections. She also assisted Gianna and Molly with data cleaning and collection, inputting the binary data for "Caitlin_played" and other manual variables. Ashley helped create early visualizations for attendance impact, cited sources in the bibliography, and ran the original linear regression and K-means analyses on the original dataset before adding the lagged variables.

Bibliography

Associated Press. “WNBA Breaks Single-Season Attendance Mark.” *ESPN*, www.espn.com/wnba/story/_/id/46040489/wnba-breaks-single-season-attendance-mark. Accessed 22 Sept. 2025.

ESPN Internet Ventures. “Caitlin Clark 2024 Stats Per Game - WNBA.” *ESPN*, https://www.espn.com/wnba/player/gamelog/_/id/4433403/type/wnba/year/2024, Accessed 17 Sept. 2025.

“Iowa Hawkeyes Women’s Basketball School History.” *Sports-Reference.com*, 2025, www.sports-reference.com/cbb/schools/iowa/women/. Accessed 17 Sept. 2025.

Judge, Lawrence W., and Jeffrey C. Petersen. "The Caitlin Clark Effect: Evidence of Athlete-Driven Market Disruption in Women’s Collegiate Basketball." *Journal of Applied Sport Management*, vol. 17, no. 2, https://trace.tennessee.edu/cgi/viewcontent.cgi?article=1661&context=jasm. Accessed 17 Sept. 2025.

McGuire, Corbin. “The Caitlin Clark Effect.” *NCAA.Org*, 15 Feb. 2024, www.ncaa.org/news/2024/2/15/media-center-the-caitlin-clark-effect.aspx. Accessed 02 Oct. 2025.

“WNBA Attendance.” *Across the Timeline*, acrossthetimeline.com/wnba/data.html. Accessed 02 Oct. 2025.

Zuritsky, Harrison. “The Caitlin Clark Effect Is an Economic Engine for Women’s Basketball.” *DCReport.Org*, 10 Jan. 2025, www.dcreport.org/2025/01/10/the-caitlin-clark-effect-is-an-economic-engine-for-womens-basketball/. Accessed 02 Oct. 2025.

Appendix

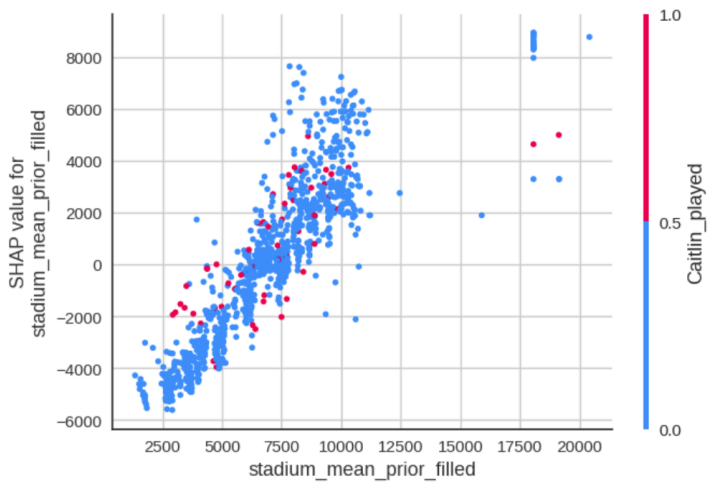

Figure 1. SHAP Dependence Plot Showing Caitlin Clark's Influence on Prior Stadium Attendance.

There is a general positive correlation between "stadium_mean_prior_filled" and its SHAP value. The color of the points represents played (red) and didn't play (blue). The red dots primarily have higher SHAP values, suggesting that her playing boosts predicted attendance.

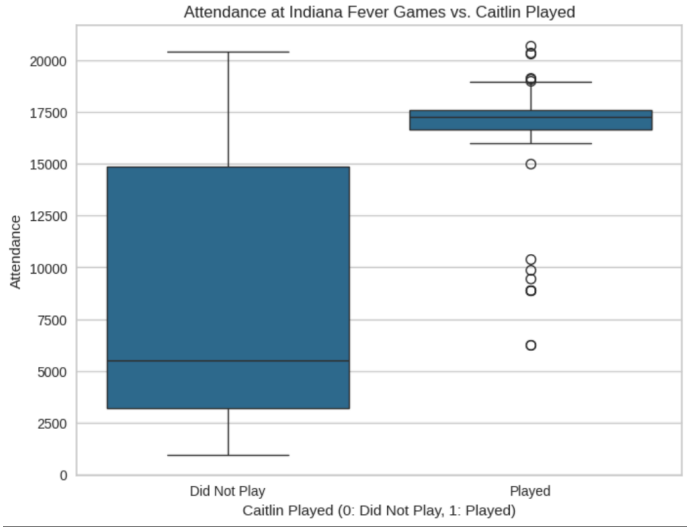

Figure 2: Comparison of Indiana Fever Attendance When Caitlin Played vs. Did Not Play

The Indiana Fever attendance significantly increases when Caitlin Clark plays vs. when she does not. There are a few outliers, but there is a huge difference in attendance numbers, favoring when she plays.

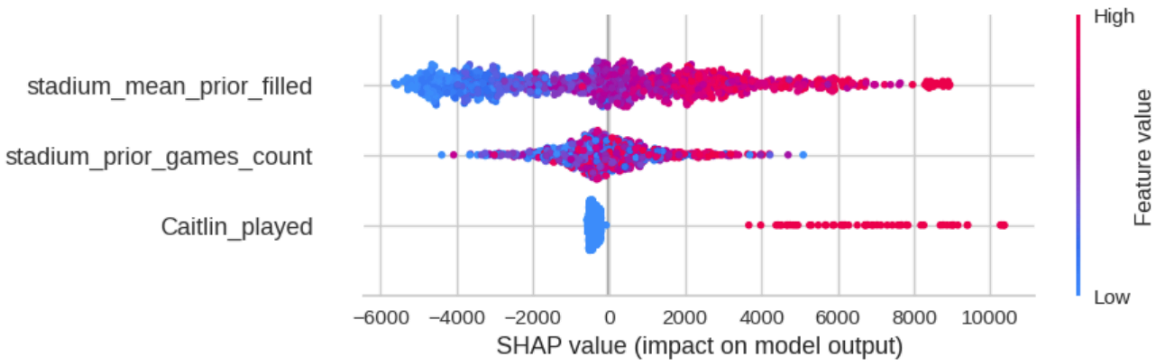

Figure 3. Feature Influence on Model Based on SHAP Values

This SHAP plot shows the top three features impacting the model. The most influential was Caitlin_played, showing her strong impact on attendance.

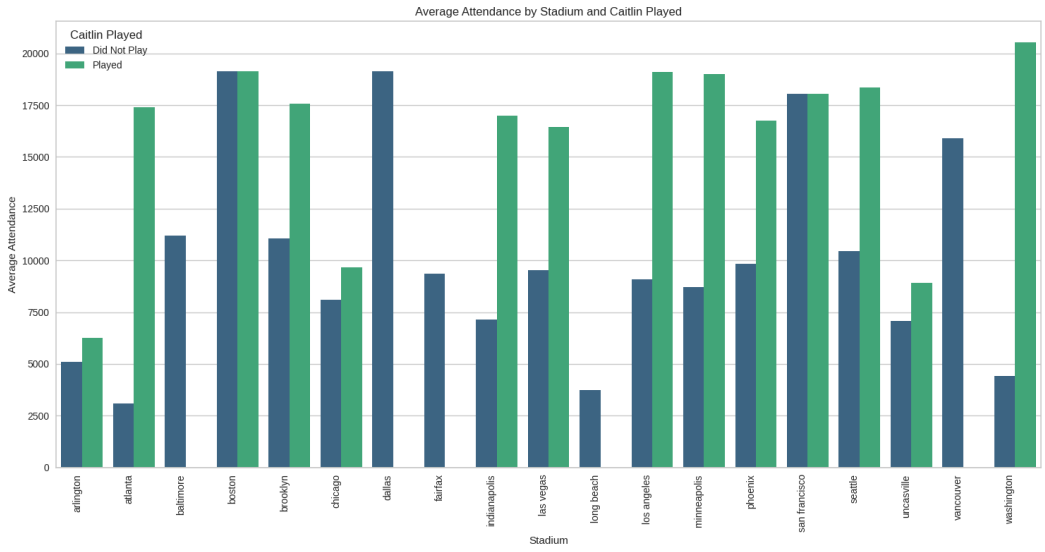

Figure 4. Caitlin Clark's Influence on Average Attendance by Stadium

This figure shows the average attendance across the cities where WNBA teams play, color-coded by whether or not Caitlin played. When Caitlin played, the stadium experienced the highest or equal-highest average attendance compared to when she did not play.

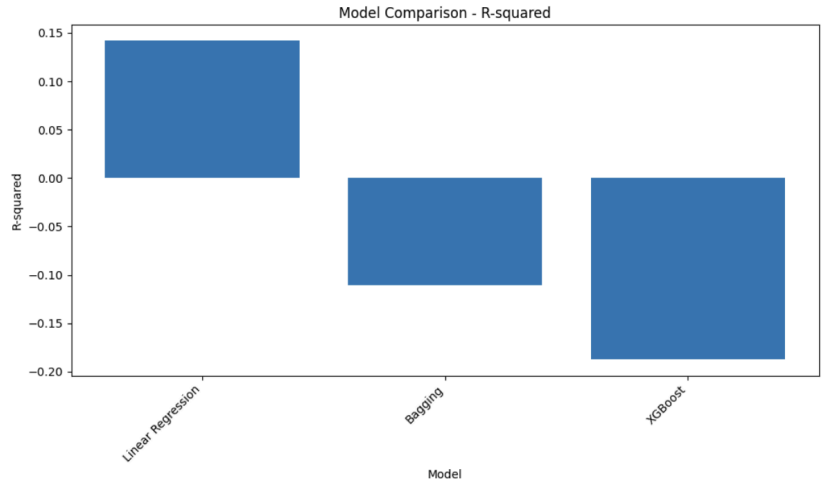

Figure 5. Model R^2 Value Comparison

The figure shows the R^2 value model comparison (linear regression, bagging, and XGBoost) before tuning occurred. This shows that linear regression has the highest R^2 value and XGBoost has the lowest.

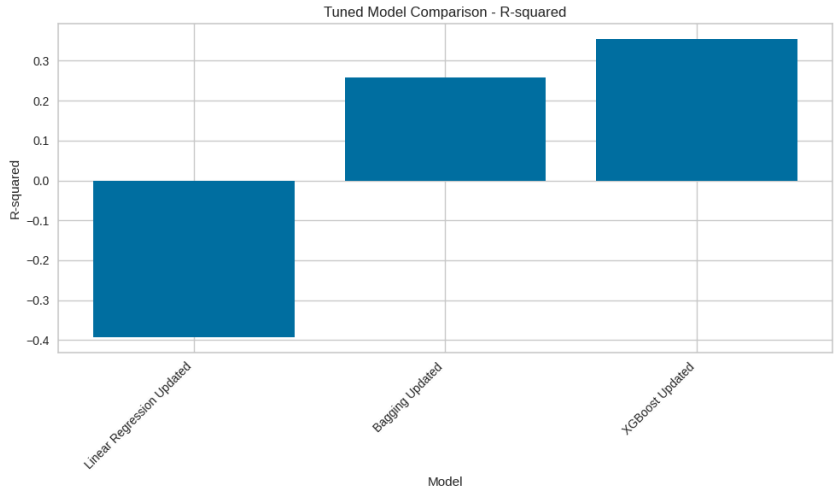

Figure 6. Tuned Model R^2 Value Comparison

The figure shows the **tuned** R^2 value model comparison (linear regression, bagging, and XGBoost) using the optimal hyperparameters. This shows that XGBoost has the highest R^2 value, which was the model we used in our analysis. Linear regression changed drastically, and now has the lowest R^2 value.

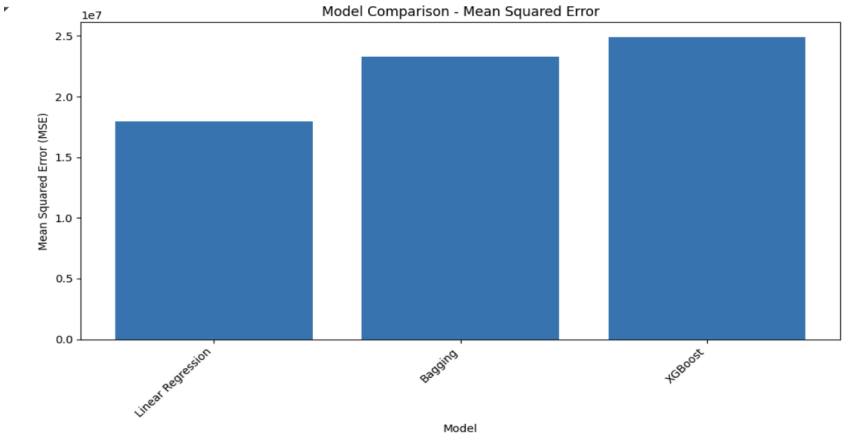

Figure 7. Mean Squared Error Model Comparison

The figure shows the mean squared error (MSE) value model comparison (linear regression, bagging, and XGBoost) before tuning occurred. This shows that XGBoost has the highest MSE and linear regression has the lowest.

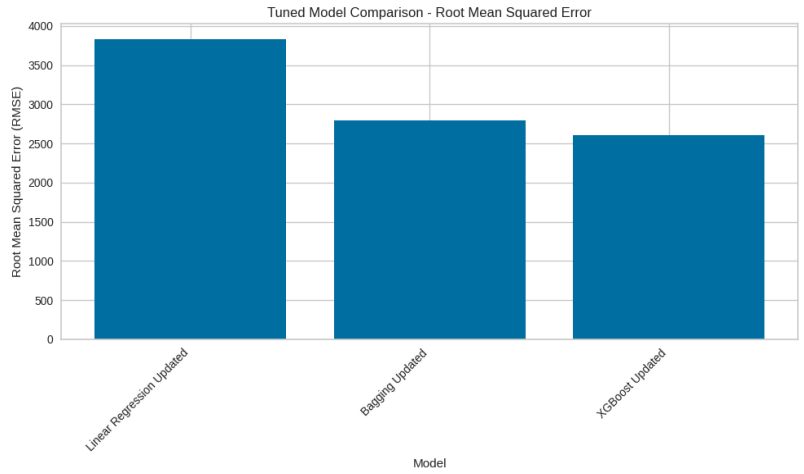

Figure 8. Tuned Root Mean Squared Error Model Comparison

The figure shows the **tuned** mean squared error (MSE) value model comparison (linear regression, bagging, and XGBoost). This shows that linear regression has the highest MSE and XGBoost has the lowest, the opposite of the previous untuned model.

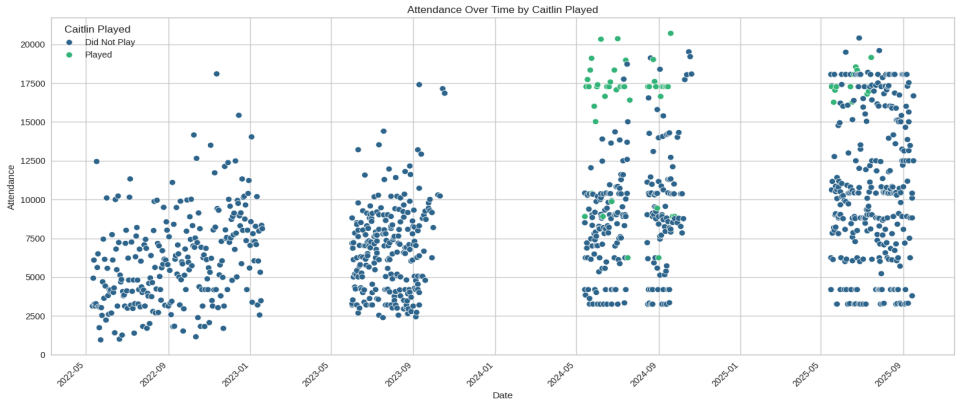

Figure 9. Clark's Impact on Attendance Over 2022-2025 Seasons

In this visualization, there is a general increase in attendance across the board in both the 2024 and 2025 (present) seasons. The green represents when Caitlin plays, and for the most part, these dots are among the highest attendance levels. However, interestingly enough, attendance overall seems to increase and become more widespread than in the 2022 and 2023 seasons.

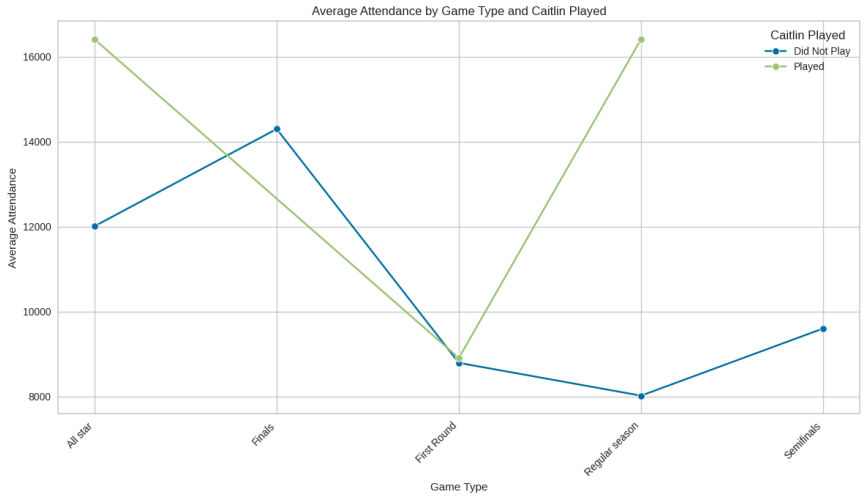

Figure 10. Line Graph of Average Attendance by Game Type When Clark Played vs. Didn't Play

This line graph compares average attendance when Clark played versus when she didn't in different game types. The highest average attendance was during the regular season when she did play. This shows the impact different game types can have on attendance.

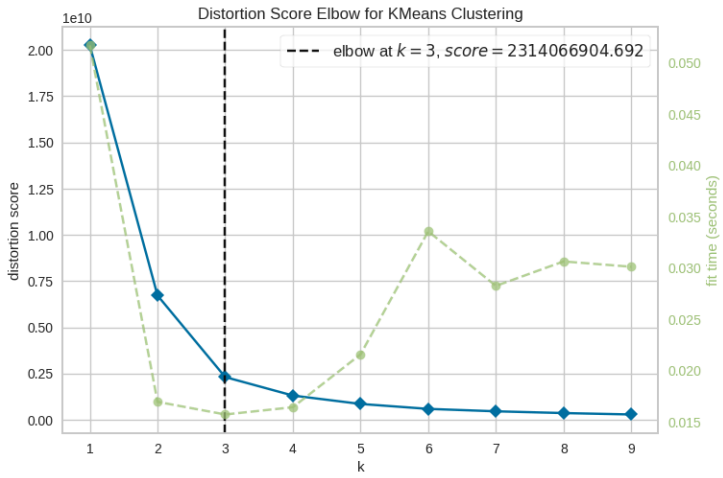

Figure 11. K-Means Clustering

A K-Means Clustering model was run. This visualization shows that 3 clusters are the ideal amount for our dataset if we used K-means clustering.

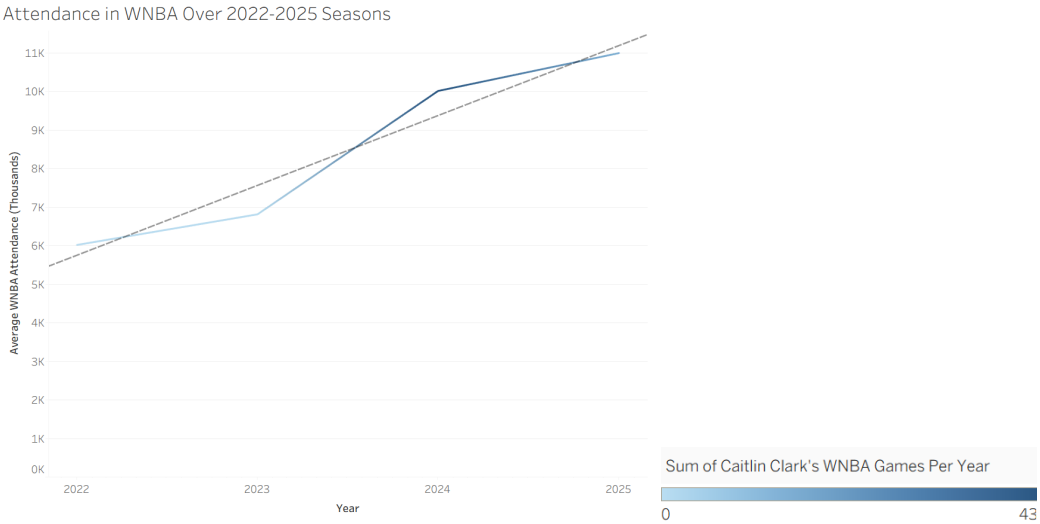

Figure 12. Attendance in WNBA over 2022-2025 Season

This line graph shows over 3 seasons how attendance has increased. The line got darker as Clark played more games. Based on that, the more games she played, the attendance steadily increased. There is also a gray trend line that displays a positive relationship between the variables.

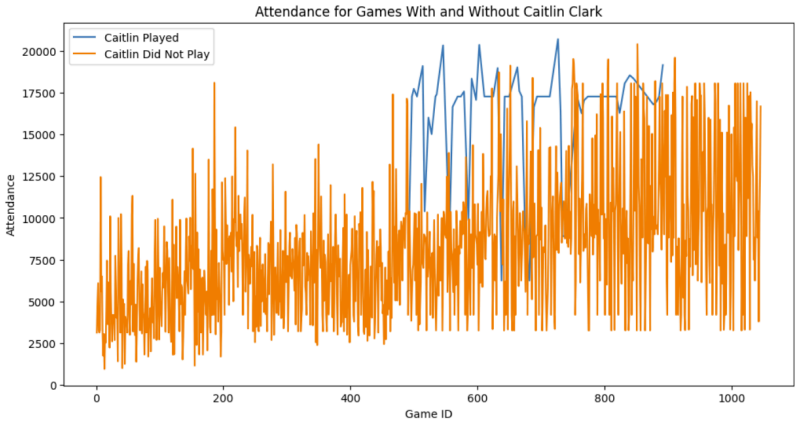

Figure 13 Line Graph for Attendance for Games With and Without Caitlin Clark (Early Visualization)

This line graph shows the attendance for WNBA games from 2022 to 2025 when Caitlin played (blue line) and did not play (orange line). The x-axis represents the Game ID. Based on the data, when Caitlin played, there was a general increase in attendance. Attendance was also steadier and more consistent when Caitlin played.

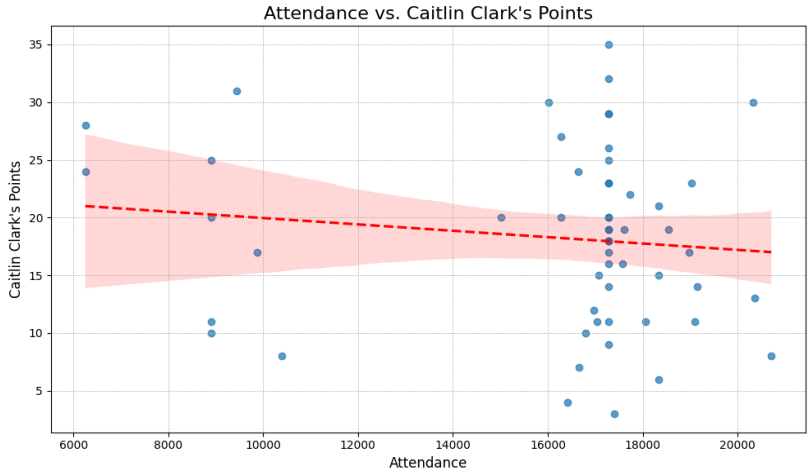

Figure 14 Attendance vs. Caitlin Clark's Points (Early Visualization)

This scatter plot shows the relationship between the number of points Caitlin Clark scored when she played and the number of people in attendance. Each plot represents a single game. The regression line suggests a slightly negative correlation, indicating that when attendance is higher, Caitlin Clark scores fewer points. One possible interpretation is that larger crowds create more pressure and may affect Clark's scoring. However, this relationship is weak and could require a deeper analysis to see if other factors influence the attendance and scoring.