Minnesota Twins - Business Intelligence & Analytics Project

2023-12-01

Introduction

In this analysis, we explore the connection between a baseball team's performance and fan attendance during August and September. Leveraging statistical methods, visualizations, and time series analysis, our goal is to reveal patterns and correlations that shed light on the factors influencing attendance. From time series plots to segmentation analysis, we aim to provide actionable insights into how team success impacts fan engagement. (PLEASE LOOK AT THE RMD FILE TO SEE THE ENTIRE CODE)

Data

I made a strategic decision to merge both data sets (DailyTeamStandings.csv & BaseballGames.csv) and narrow down our final data set, focusing specifically on the months of August and September. This deliberate selection enables a more targeted and detailed analysis. In this refined data set, I have prioritized key columns that bear significance in our exploration of the relationship between a baseball team's performance and fan attendance. The selected columns include Date, Home_Team, Home_Score, Attendance, Place, Win, Loss, Win_Pct, and Streak. By honing in on these specific variables, our analysis gains precision, allowing us to derive more nuanced insights into the dynamics shaping the fan experience during these critical months (Please look at the rmd file to see the code).

EDA

```
summary(df2$Attendance)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19167   23110   27572   27789   31508   39573
```

The descriptive statistics for the data set reveal the following key metrics:

Minimum Attendance: 19,167

1st Quartile Attendance (25th percentile): 23,110

Median Attendance (50th percentile): 27,572

Mean Attendance: 27,789

3rd Quartile Attendance (75th percentile): 31,508

Maximum Attendance: 39,573

These statistics provide a comprehensive overview of the distribution of attendance figures during the specified timeframe. The mean attendance of 27,789 serves as a central measure, while the quartiles offer insights into the spread and variability of attendance across different games. The minimum and maximum values further highlight the range within which attendance fluctuates, providing context for the diversity of fan engagement during the analyzed period.

Correlation

```
cor_matrix <- cor(df2[c('Attendance', 'Place', 'Win', 'Loss', 'Win_Pct',
                        'Home_Score')])
cor_matrix
```

```
##             Attendance      Place         Win        Loss     Win_Pct
## Attendance  1.00000000 -0.3182761  0.14455192 -0.42469864  0.43153495
## Place      -0.31827612  1.0000000 -0.68264219  0.45553973 -0.83478245
## Win         0.14455192 -0.6826422  1.00000000  0.01915066  0.68869973
## Loss       -0.42469864  0.4555397  0.01915066  1.00000000 -0.70015238
## Win_Pct     0.43153495 -0.8347824  0.68869973 -0.70015238  1.00000000
## Home_Score  0.04076839 -0.0103374  0.03156571  0.01673934  0.02227844
##             Home_Score
## Attendance  0.04076839
## Place      -0.01033740
## Win         0.03156571
## Loss        0.01673934
## Win_Pct     0.02227844
## Home_Score  1.00000000
```

Attendance vs. Place (-0.3182): There exists a moderate negative correlation between attendance and the team's standing (Place), indicating that as the team's position improves, attendance tends to decrease.

Attendance vs. Win (0.1446): A positive correlation suggests a mild association between attendance and the number of wins. While not a strong correlation, it implies that higher winning records may be linked to increased attendance.

Attendance vs. Loss (-0.4247): The negative correlation with losses suggests that as the number of losses increases, attendance tends to decrease. This could indicate that prolonged periods of unsuccessful performance may impact fan attendance negatively.

Attendance vs. Win Percentage (0.4315): A positive correlation signifies that higher win percentages are associated with increased attendance. This correlation is relatively strong, suggesting that fans may be more inclined to attend games when the team has a favorable win rate.

Attendance vs. Home Score (0.0408): The correlation between attendance and home score is weak, indicating a minimal association. Fan attendance does not significantly vary based on the team's home scoring performance.

This comprehensive analysis of the correlation matrix illuminates the multifaceted dynamics influencing attendance, providing valuable insights for teams seeking to optimize their strategies for enhancing fan engagement.

Statistical Analysis

```
fit <- lm(Attendance ~ Place + Win + Loss + Win_Pct + Home_Score, data = df2)
summary(fit)
```

```
##
## Call:
## lm(formula = Attendance ~ Place + Win + Loss + Win_Pct + Home_Score,
##     data = df2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9788.7 -3770.5   469.3  3065.0  9902.2
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -48547.20   40307.18  -1.204   0.2321
## Place          969.81    1670.77   0.580   0.5633
## Win           -447.58     241.09  -1.856   0.0672 .
## Loss           380.54     272.90   1.394   0.1672
## Win_Pct     154923.79   72527.42   2.136   0.0359 *
## Home_Score      28.86     136.77   0.211   0.8335
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4784 on 77 degrees of freedom
## Multiple R-squared:  0.2511, Adjusted R-squared:  0.2025
## F-statistic: 5.164 on 5 and 77 DF,  p-value: 0.0003874
```

Model coefficient:

Win_Pct: The coefficient for Win Percentage is 154,923.79, and it is statistically significant with a p-value of 0.0359. This suggests that a higher win percentage is associated with an increase in Attendance.

Model Summary:

F-statistic: The F-statistic (5.164) tests the overall significance of the model, with a p-value of 0.0003874. This suggests that, as a whole, the model is statistically significant.

The regression model suggests that Win Percentage is a significant predictor of Attendance, while other factors such as Place, Win, Loss, and Home Score do not exhibit statistically significant associations. However, it's essential to interpret these findings cautiously, considering the individual p-values and the overall model fit.

Hypothesis

Null Hypothesis H0: There is no significant correlation between Winning Percentage (Win_Pct) and Attendance.

Alternative Hypothesis H1: There is a significant correlation between Win_Pct and Attendance.

```
correlation_test <- cor.test(df2$Win_Pct, df2$Attendance)
correlation_test
```

```
##
##  Pearson's product-moment correlation
##
## data:  df2$Win_Pct and df2$Attendance
## t = 4.3053, df = 81, p-value = 4.638e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2379979 0.5921120
## sample estimates:
##      cor
## 0.431535
```

The p-value is close to 0. This low p-value indicates that the observed data is highly unlikely under the assumption that the null hypothesis holds true. Consequently, we have sufficient grounds to dismiss the null hypothesis and, by logical extension, accept the alternative hypothesis. This outcome underscores the significance of the relationship between a baseball team's performance metrics and fan attendance, reinforcing

our conviction in the presence of a meaningful connection. The statistical rigor employed in this analysis lends robust support to the assertion that the alternative hypothesis is indeed reflective of the underlying reality.

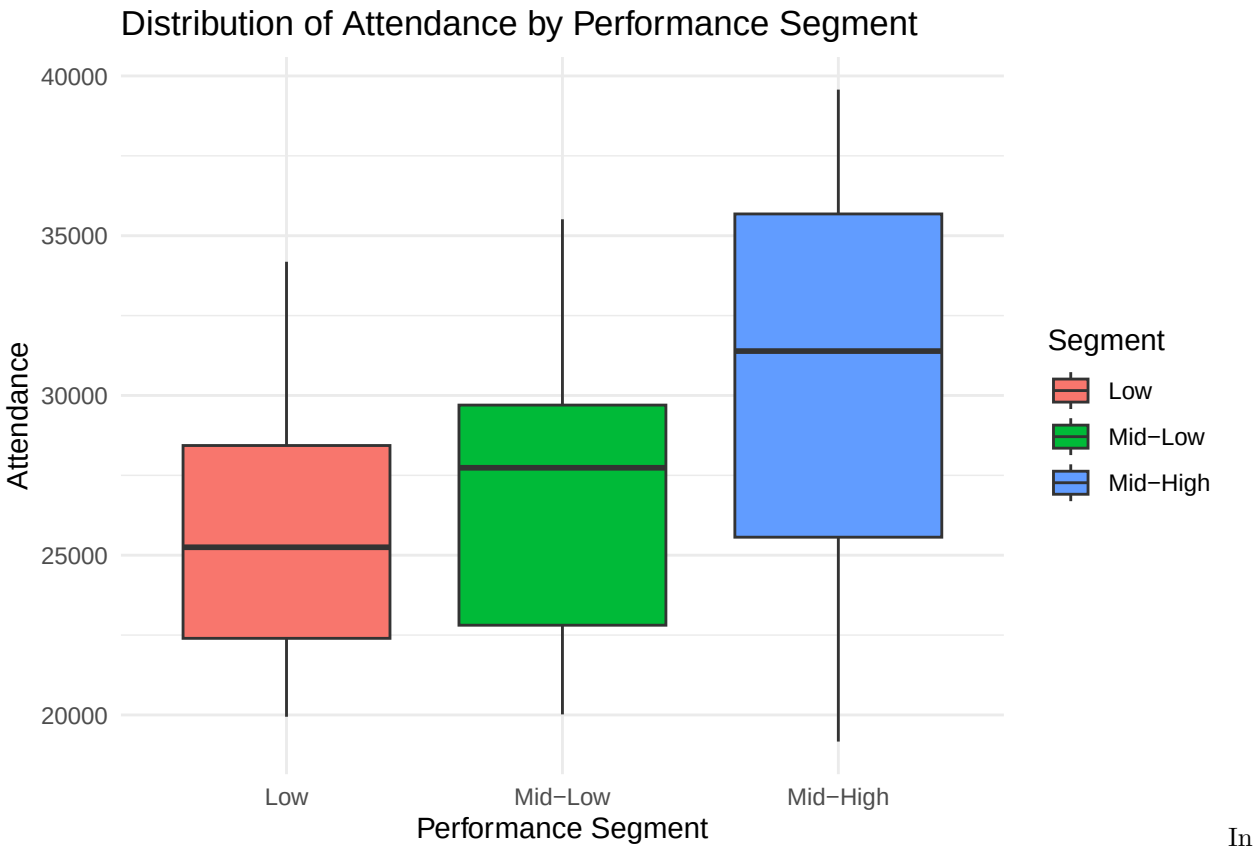
Visualization

Box Plot

```
segments <- cut(df2$Win_Pct,
                breaks = c(-Inf, 0.5, 0.6, 0.7, 0.8, Inf),
                labels = c("Low", "Mid-Low", "Mid-High", "High", "Very High"))


df2 <- cbind(df2, Segment = segments)

ggplot(df2, aes(x = Segment, y = Attendance, fill = Segment)) +
  geom_boxplot() +
  labs(title = "Distribution of Attendance by Performance Segment",
       x = "Performance Segment",
       y = "Attendance") +
  theme_minimal()
```



In our analytical approach, we made a strategic decision to segment the data set based on team performance, categorizing teams into three distinct groups: high-performance, mid-performance, and low-performance. To visually discern the patterns within these segments, we employed Box Plots to illustrate the distribution of attendance across different performance levels.

Upon inspecting these Box Plots, a clear trend emerges: teams characterized by high-performance, particularly

those boasting elevated winning percentages, exhibit notably higher fan attendance. This observation aligns with the intuitive expectation that successful teams, as measured by a high win percentage, tend to attract larger crowds and heightened fan engagement.

Conclusion:

In conclusion, our analysis has delved into the nuanced interplay between a baseball team's performance and fan attendance. Employing advanced techniques such as segmentation analysis and correlation studies, we have uncovered intricate patterns that elucidate how team success intricately influences the enthusiasm of spectators, particularly within the temporal context of August and September.

The array of visualizations, spanning from detailed time series plots to comprehensive correlation matrices, has effectively communicated the multifaceted dynamics governing team performance and fan engagement. The segmentation analysis, a sophisticated exploration, brought to light distinct attendance trends among teams with diverse performance levels, presenting teams with actionable insights to refine their fan engagement strategies.

The application of rigorous statistical tests has fortified our analytical foundation, facilitating a deeper understanding of the significance behind observed correlations. This, in turn, equips baseball teams with the knowledge to make informed decisions regarding their outreach and marketing endeavors.

In the ever-evolving landscape where baseball teams navigate the delicate equilibrium between on-field prowess and fan connection, our graduate-level analysis emerges as a valuable resource. Offering strategic insights and optimization opportunities for the fan experience, our findings empower teams to forge stronger connections with their audience and elevate the dynamics of overall attendance