Baseball Advertising Analysis

Matt

2024-08-06

Setting up my Environment

```
library(tidyverse)
```

Loading the Data

```
teams <- read.csv("team.csv")
```

Cleaning the Data

```
cleaned_teams <- teams %>%
  filter(year >= 2010) %>%
  mutate(avg_attendance = attendance/home_games) %>%
  select(-"attendance")
```

Checking Correlations

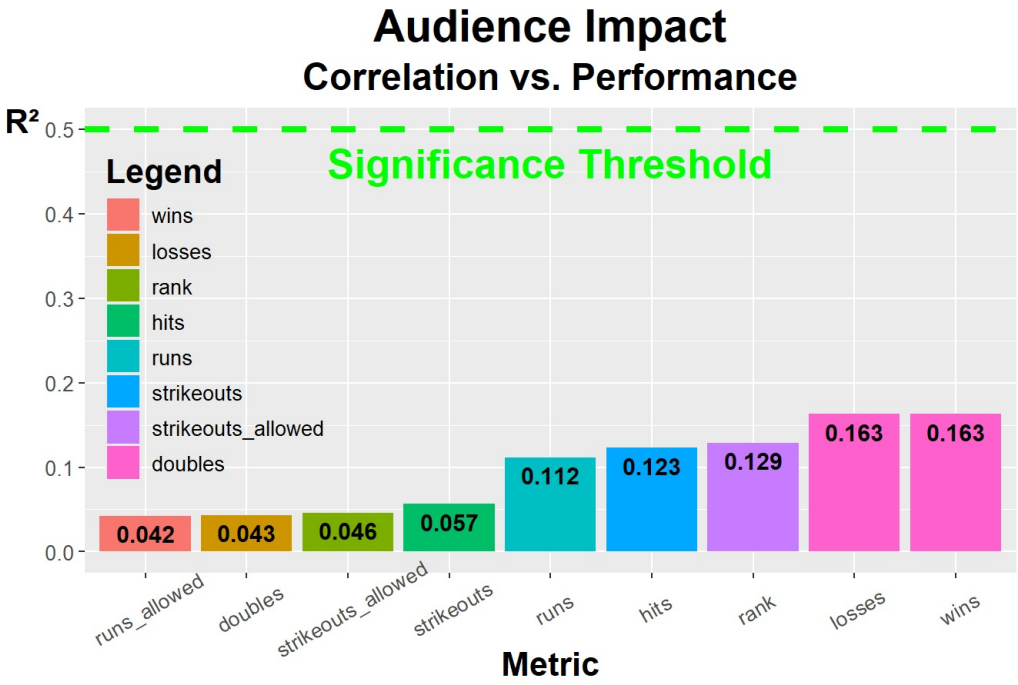
```
for(x in 1:length(colnames(cleaned_teams))-1){
  if(is.numeric(cleaned_teams[1,colnames(cleaned_teams)[as.numeric(x)]])){
    i <- colnames(cleaned_teams)[as.numeric(x)]
    j <- round(cor(cleaned_teams$avg_attendance,cleaned_teams[colnames(cleaned_teams)[as.numeric(x)]]),3)
    kvp <- c(i,j)
    if(x == 1){
      df <- data.frame(Variable = c(i), Correlation = c(j))
    }
    else{
      df <- rbind(df,kvp)
    }
  }
}
arrange(df,desc(abs(as.numeric(Correlation))))
```

```
##                  Variable Correlation
## 1                    wins       0.404
## 2                  losses      -0.404
## 3                    rank      -0.359
## 4                    hits       0.351
## 5                    runs       0.334
## 6              strikeouts      -0.238
## 7      strikeouts_allowed       0.215
## 8                 doubles       0.208
## 9            runs_allowed      -0.206
## 10          walks_allowed      -0.199
## 11            earned_runs      -0.196
## 12    earned_runs_average      -0.196
## 13                shutout       0.181
## 14                at_bats       0.175
## 15                  saves       0.175
## 16           stolen_bases      -0.169
## 17        caught_stealing      -0.169
## 18 bat_performance_factor       0.157
## 19           hit_by_pitch       0.153
## 20           hits_allowed      -0.146
## 21              home_runs       0.137
## 22      home_runs_allowed      -0.134
## 23                 errors      -0.134
## 24    fielding_percentage       0.134
## 25         complete_games       0.128
## 26           outs_pitched       0.095
## 27             home_games       0.094
## 28                  walks       0.093
## 29        sacrafice_flies       0.091
## 30   pitching_park_factor       0.087
## 31                triples       0.061
## 32           double_plays      -0.047
## 33                  games       0.026
## 34                   year       0.011
```

Visualizing Correlations

```
plot_df <- df %>%
  mutate(r_squared = as.numeric(Correlation)^2) %>%
  filter(r_squared >= 0.04) %>%
  arrange(desc(r_squared))

ggplot(data=plot_df) +
geom_col(mapping = aes(x = reorder(Variable, r_squared), y = r_squared, group = factor(r_squared), fill = factor(
r_squared))) +
geom_text(mapping = aes(x = reorder(Variable, r_squared), y = r_squared, label = round(r_squared, digits = 3),  f
ontface = "bold"), size = 4, vjust = 1.5) +
theme(axis.text.x = element_text( vjust = 0.8, angle = 30, size = 10),
      axis.text.y = element_text(size = 10),
      axis.title.y = element_text(angle = 0, size = 16),
      axis.title.x = element_text(vjust = 7, size = 16),
      title = element_text(size = 18, face = "bold"),
      plot.title = element_text(hjust = 0.5),
      plot.subtitle = element_text(hjust = 0.5),
      legend.position = "inside",
      legend.position.inside = c(.14, .55),
      legend.background = element_blank(),
      legend.title = element_text(size = 16),
      legend.text = element_text(size = 10)) +
labs(title = "Audience Impact", subtitle = "Correlation vs. Performance", fill = "Legend", x = "Metric", y = "R\u
00B2") +
scale_fill_discrete(labels = plot_df[['Variable']]) +
geom_hline(yintercept = 0.5, linetype = "dashed", linewidth = 1.5, color = "green") +
annotate("text", x = 5, y = 0.46, label = "Significance Threshold", color = "green", fontface = "bold", size = 7)
```



Takeaway

Statistically significant r-squared values are typically above 0.5, however none of the game statistics seem to indicate any significant correlation to average park attendance.