Deliverables, Part 1

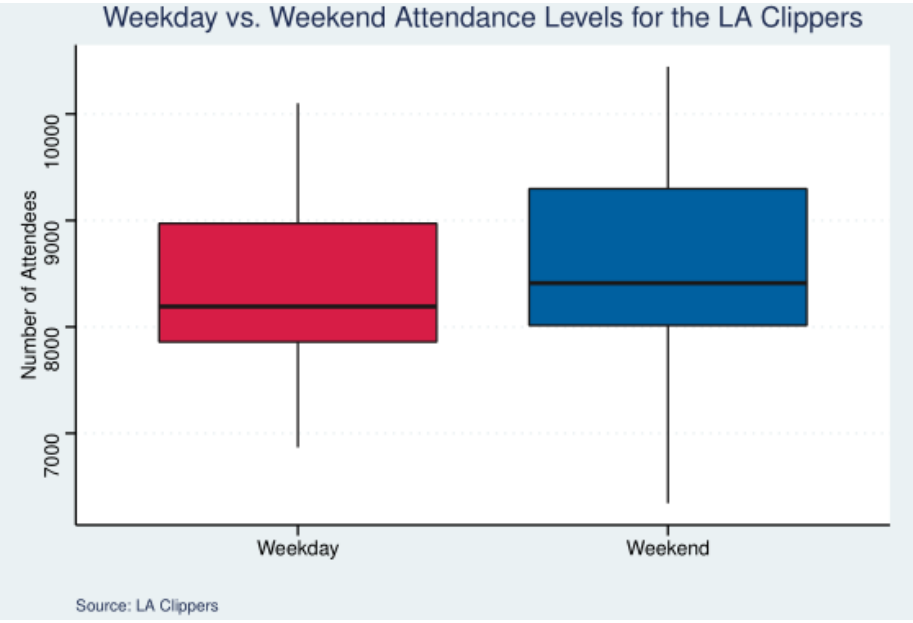
Linus Jen

*To follow along with the results shown here, please open up the "Deliverables Part 1.pdf" that shows the step-by-step code and results as to how I found these answers. Each question on this sheet has its own respective header and code portion in the pdf, so you can follow my logic and notes as to how I tackled each problem.

Question 1: Do weekend games (Friday, Saturday, and Sunday), on average, have a higher attendance than weekday games (Monday through Thursday)?

No, the average number of fans that attended during the weekends was not statistically larger than the average number of fans that attended weekday games. This was shown in several ways. First, I ran a t-test to see if the difference in average attendance was significant. Not only were the means off by an average of 300 people (8606 on weekends vs. 8323 on weekdays), but the t-test came back with a small t-score of -.832 and a large p-value of .21. In addition to this, the boxplot (shown below) shows that the distribution between average weekday and weekend attendance are almost exactly the same overall, with the weekend distribution being slightly wider.

In terms of how I coded this up, I first checked the dates to see whether they were weekday or weekend games. I used this label to group them, and found the mean for both groups. A t-test was run with its respective function, and a simple boxplot was created to compare the distributions. While weekend games seem to have larger variance in terms of the number of people in attendance, overall they seem to be very similar.



Question 2: Identify and rank the top 4 opponents with the highest average number of attendances.

Opponent	Average Attendees
Boston Celtics	10101
Golden State Warriors	9798
Los Angeles Lakers	9477
New York Knicks	9432

These results don't seem to be quite shocking. Notice that three of the four teams have a long history in the NBA, as the Celtics, Knicks, and Lakers make up some of the winningest teams in the NBA. Following this, all four teams come from large metropolitan cities, and thus have the capability of drawing in more fans growing up. Thus, more people would be interested in watching these teams live.

Coding wise, I used the cleaned up ticket_scan_data dataset, grouped by the game_number and Opponent columns, and first found the number of fans in attendance per game. Then, I grouped by the opponent, and created a summary table based off the average number of fans in attendance for each game against a specific team. The full of the teams played and the average attendees for each team can be found in the "Deliverables Part 1 Code.pdf".

Question 3: Identify and rank the top 10 sections that are, on average, the most filled to their capacity.

Section Name	Average Fill Percentage
207	.794
309	.790
106	.786
118	.782
215	.774
116	.773
206	.766
105	.764
310	.758
217	.753

With the seating chart provided, I discovered that nine of the ten sections on this list are either directly behind one of the backboards, or on the higher levels behind the backboards. The only section shown above that is not behind one of the backboards is section 118, which is very near to the Clippers's locker room. From my work with UCLA Athletics, my experience is that these spots tend to be cheaper than other seats, and the backboards "impede" one's view of the

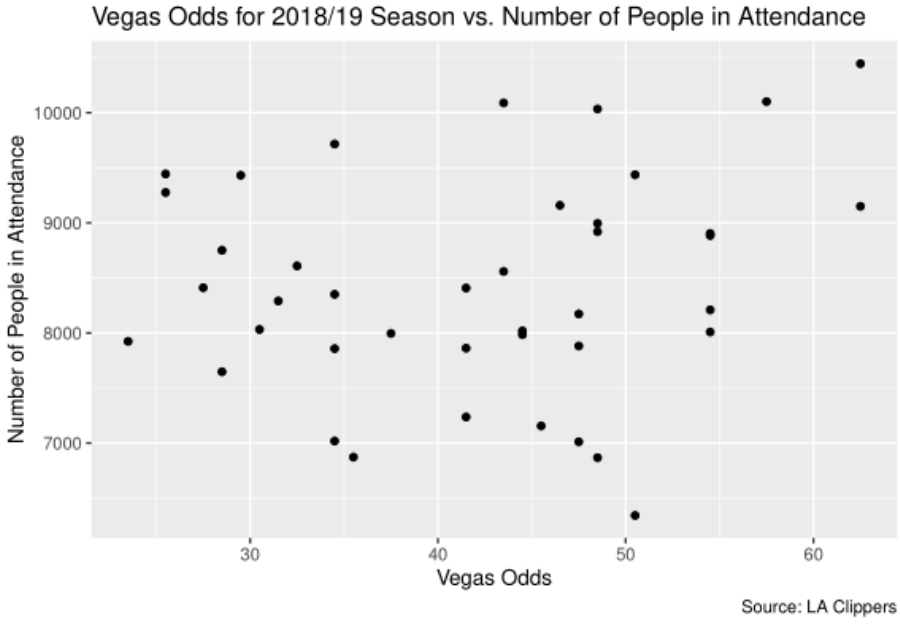
game, and thus are not very popular seats. However, because of the cheaper prices, the Clippers may actually have more success filling up these spots, but more investigation will be needed to prove that this is the case. Note that not all the section data was given, as sections 101 and 102 seem to be unoccupied or not included in the dataset, as were all the court sections (marked with CT).

To solve this problem, I first used the ticket_scan_data to group by the section name and game, and then counted how many seats were filled per section and game. Then, I joined this dataset with the seating_chart data, using the column of total available seats per section to find the percentage of seats occupied per section of each game. Then, I found the average fill percent for each section, and sorted it into the summary table, shown above.

Question 4: Is there a correlation between the opponent teams having a higher Vegas Odds Score (indicating a higher probability of winning a championship) and higher attendance?

Yes, there seems to be a minor correlation between higher Vegas Odds Scores and higher attendance. Three methods were used to find the correlation, but the highest correlation found that these two variables had a value of .375.

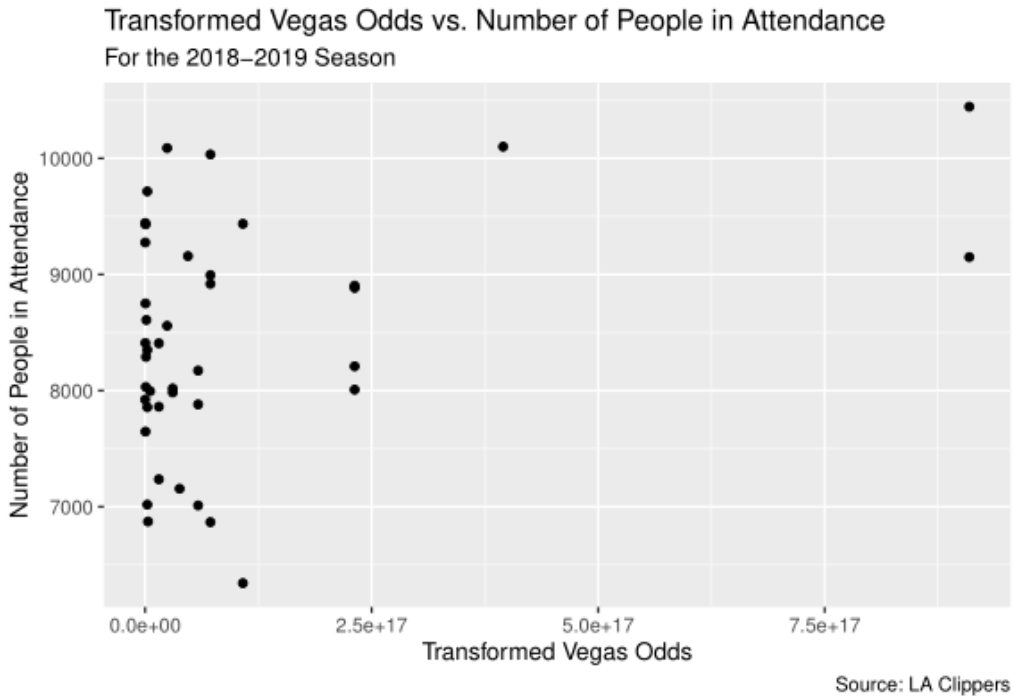
The first method was to simply find the correlation between the Vegas Odds and the number of attendees per game. The Vegas Odds Score data was found in the team_data, and was joined to the average number of attendees per game data frame created in a different section. From this, an R value of .161 was obtained – a decent baseline. A plot of this data, using Vegas Odds to predict attendance, can be seen below.



The second method used was to use the powerTransform() function, which finds the optimal power for the Vegas Odds such that the new, transformed values would mirror a normal

distribution. From this method, a lambda value of 1 was given, meaning that the data did not need to be transformed. Thus, no improvement in the R value was found here.

Lastly, I used the inverse response plot to find the optimal power to transform the number of attendees to have a distribution similar to that of the Vegas Odds Scores. This yielded the most success of all, returning an R value of .375. A scatterplot of the attendance level vs. transformed Vegas Odds can be found below.

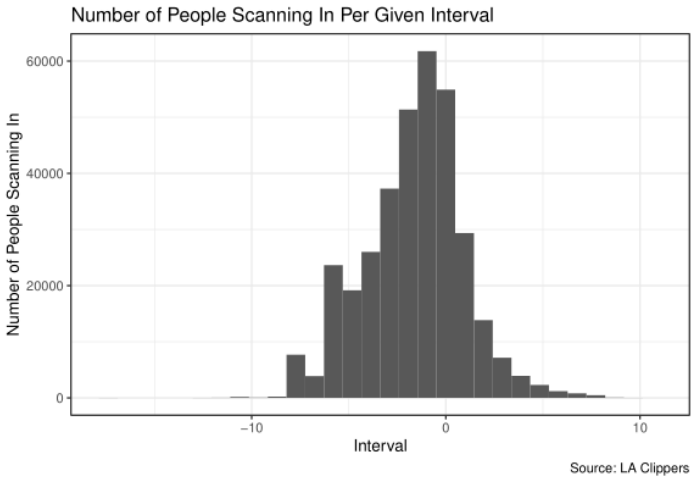


Again, the code and logic behind these steps can be found in the "Deliverables Part 1 Code.pdf" file.

Question 5: On average, which 15-minute period before or after the start of game has the highest number of people scanning?

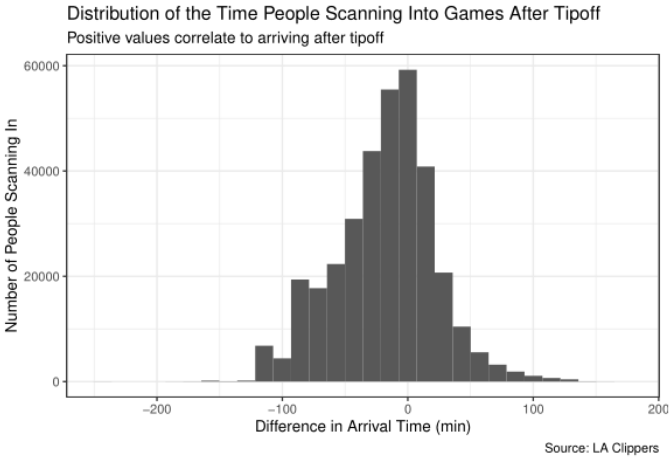
The interval from 30 minutes prior to tipoff to 15 minutes before the game starts tends to see the greatest amount of foot traffic going into the Staples Center. This can be clearly seen in the histograms below, as the value near -1 is the peak for the number of people scanning into games. A table of the top 5 busiest intervals can be found below.

Interval Number	Time Interval	Average Number of Scans
-1	30 – 15 minutes before tipoff	1507
0	15 – 0 minutes before tipoff	1340
-2	45 – 30 minutes before tipoff	1253
-3	60 – 30 minutes before tipoff	909
1	0 – 15 minutes after tipoff	716



The figure to the left shows the distribution when people would scan in, based off the interval before or after the game had started. Positive values correlate to after tipoff, while negative values correlate to arriving early. Interval values were attained by dividing the difference in scan times and tipoff, and then rounding these values down.

The figure to the right shows the distribution of when people would scan in. Negative values correlate to arriving early, while positive values correlate to arriving after tipoff. Similar to the histogram above, we see that most people arrive right before the game starts, which agrees with our findings that the interval from 30 minutes to 15 minutes before the start of the game sees the most people scanning in.



A short summary on how this was coded up: In the data clean-up file, we had already converted all the times for the games to be in Pacific Standard Time, so the first step to do was subtract the scan time for each individual with the event start time. Then, we grouped them into intervals by dividing the difference in times (in minutes) by 15, and rounding all the values down. For example, a person scanning into the game at 7:20PM for a 7:30PM game would have a difference of -10, and after dividing by 15 and rounding down, would get a value of 0, representing that they came during the interval between 15 minutes before the game and tipoff. On the other hand, if the person came at 7:20PM for a 7:00PM game, the difference in times would be 20, and when divided by 15 and rounded down, we would get a value of 1. Then, a summary table adding up the total number of attendees per interval and game was created. Lastly, another summary table was made from this to find the average number of scans per interval, and sorted. A shortened table is shown above.