Project 2

Are all "Active Moviegoers" Created Equal?

Examining the Heterogeneity Amongst Self-Proclaimed Cinema

Enthusiast through Inferential Models.

April 14, 2024

1 Executive Summary

Understanding the consumer population is a pivotal aspect to releasing a successful box office film. By gaining insight into the demographic segments that the film will likely reach and how each segment may respond to the release can reveal the target audience. This report will provide modeling methods to outline the various different segments of the misleadingly, uniform term "active moviegoers" and how the underlying population propensity of viewing Wonka actually differs.

1.1 Opportunity

The underlying behavior of "active moviegoers" at the individual level can be described as continuous timing. Although a Beta-Geometric model could be argued to potentially describe these behaviors, it's more appropriate to describe the decision to watch the movie on a continuous time scale where individuals decide and act at an exact period and plunge into action rather than discrete opportunities until they make the decision. Given the usual time-varying effects and heterogeneous nature of the population, the Weibull distribution and Burr XII models will be used to provide initial intuition regarding the behavior of the population, and the latent class models built with individual Weibull distributions will provide further insight into the specifics of each sub-population.

1.2 Solution

To emphasize once again, understanding the underlying behavior of the consumer population is crucial in creating successful films. One key aspect to this success is creating films and marketing strategies to reach the appropriate, targeted audience rather than every "moviegoer" on the planet. With only about 33% of the 50,000 "active moviegoer" population watching the Wonka film within the first 3

months of release, it's clear that "active moviegoers" does not mean "will watch every film released." Of that 33%, nearly 36.44% could be classified as strictly "Wonka fanatics" and "Christmas moviegoers," which we'll call the "Christmas fanatics." Along with this fact, the models reveal that the hardcore never-buyer (HCNB) population is approximately 66% of the total population, indicating that among the remaining 67% of "active moviegoers," if they haven't watched within three months of release then they likely never will.

As I'll explore in further detail below, these summarized results ultimately reveal that nearly 36.44% of the revenue appears to be directly linked with the presence of Christmas break whereas the other 63.56% may actual resemble the true population behavior that most "active moviegoers" may share. Deciphering the key differences between these groups may reveal the true propensity of each to provide insightful information regarding future film-making and the associated marketing decisions.

2 Feature Selection/Creation

In order to build a inferential model aimed to properly understand the population propensity and their varying response to certain events, I ensured that the features I used were properly descriptive enough to fit a sufficient model. Similarly, I was equally cautious to not include too many descriptive variables that would completely undermine the robustness of the model. The features included can be found in the table below:

Feature	Definition
Release Day	Binary flag for release date of film
Friday	Binary flag for Fridays
Saturday	Binary flag for Saturdays
Sunday	Binary flag for Sundays
Holiday	Binary flag for holidays
Christmas/NY Eve	Binary flag for Christmas and New Years Eve
Stand. Trend	Standardized scaled searches for "Timothée Chalamet"

Table 1: Feature name and definitions

Without diving into the results of the obviously weak baseline Burr XII, the model highlights the importance of covariate inclusion and creation to control for certain behavioral patterns and influential events of buying habits to properly describe the underlying population propensity.

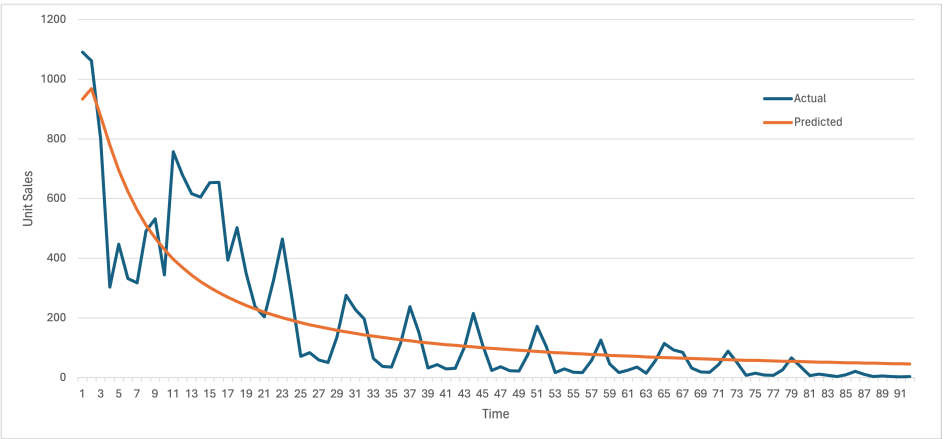

Figure 1: Baseline Burr XII

Notice the creation of individual binary flags for Friday, Saturday, and Sunday separately rather than a equal inclusion as a single feature. As seen in the plot below, there seems to be an obvious spike in ticket sales on these three days but the spike appears to vary in magnitude. Specifically, Saturday boasts much higher sales than Friday and Sunday, therefore, treating each day as their own feature would allow the model to appropriately calibrate how each day changes in sale respective to the multiplicative declining trend that can be seen. This addition was ultimately able to model the troughs and peaks much better.

Similarly, the release date and holidays appear to have large effects on the ticket sales. The inclusion of scaled searches serves the purpose of revealing behavioral differences between the subgroups rather than to improve model fitment. Perhaps this might reveal whether some groups were more or less influenced to watch Wonka due to the popularity of the actor. Additionally, it's important to recognize the potential endogenous nature of scaled searches as an increase in scaled searches for Timothée Chalamet can be argued to directly effect the unit sales of Wonka as well.

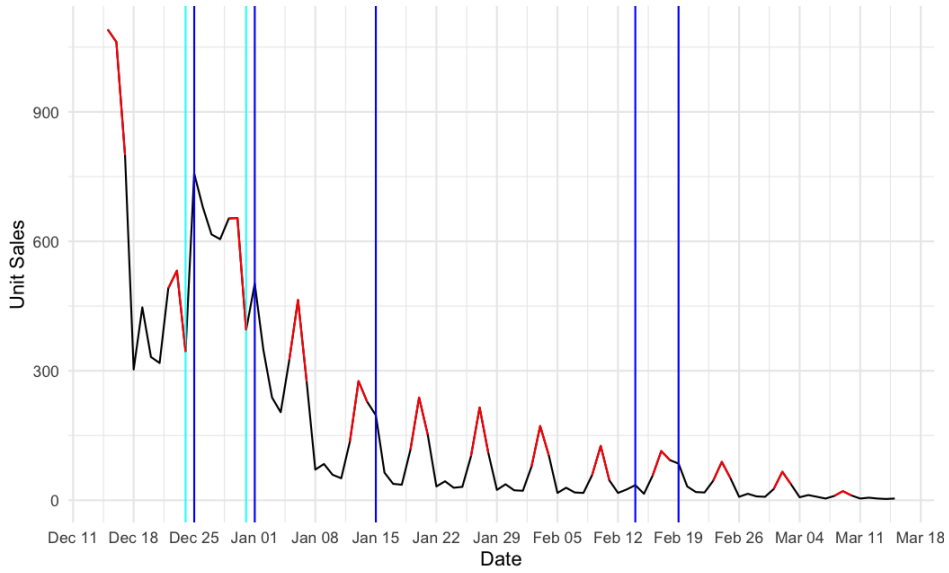

Figure 2: Plot of Unit Sales vs. Date. Red indicates Friday, Saturday, Sunday periods, blue represents holidays, and cyan represents Christmas and New Years Eve.

3 Burr XII and Weibull Model

To understand the initial propensity of moviegoers and reveal any underlying population characteristics, the Burr XII model was utilized to account for any heterogeneity and time-varying factors that moviegoers are expected to have. Below is the plot, parameters, and coefficients:

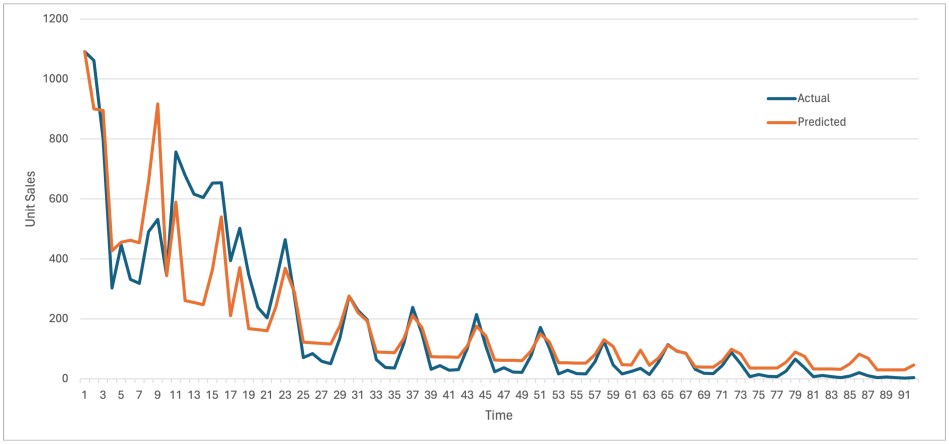

Figure 3: Burr XII + Covariates In-sample Expected vs. Predicted

Parameter	Value
r	0.0661
α	15.1707
c	1.8658
λ	0.0044

Coefficient	Value
β_{released}	1.3480
β_{friday}	0.4463
β_{saturday}	0.9564
β_{sunday}	0.8091
β_{holiday}	0.7355
β_{eve}	-0.7022

Metric	Performance
Log-Likelihood	-96217
BIC	192532
MAPE	1.3500

Table 2: Burr XII + Covariates Coefficients/Parameter/Performance

Visual inspection of the baseline Burr XII reveals the difficulty in capturing the full fluctuating wave-like nature after New Years and the period of increased sales between Christmas and New Years. This is reinforced by a MAPE of 135% indicating that there's much room for improvement. Additionally, the r-value of 0.0661 reflects an extremely heterogeneous population. Therefore, the later developed models will incorporate a spike at 0 to account for HCNB. The time-varying parameter c of 1.8658 reveals the presence of time-dependence for moviegoers. According to this model, moviegoers are increasingly likely to watch the movie as the time progresses given they didn't watch in the previous period due to the "stretching" of time.

The poor MAPE indicates that there might be issues with the overwhelming count of HCNB and the obvious hump near Christmas. Additionally, as you can find below, the exclusion of HCNB leads to an an R-value of 15577.4808, indicating an absence of heterogeneity in the observed buyers. Subsequently, the extreme homogeneity of the spiked Burr XII reduces into a spiked Weibull model that saves an extra parameter. This fact is validated through the value of r/α from the Burr XII being equal to the optimal λ of the Weibull, which we can see are equivalent below. You will notice that the decision to reduce the Burr XII to a Weibull yields identical results while reducing the BIC by 10. The entire results, plots, and analysis of the Weibull model can be found below:

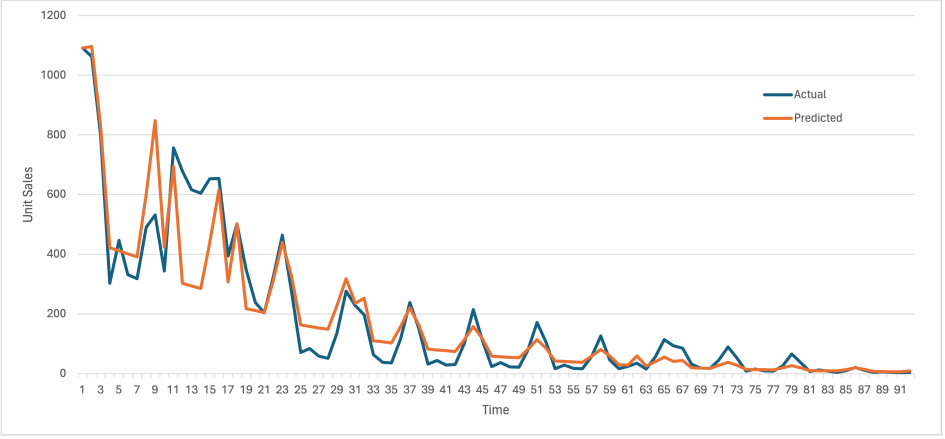

Figure 4: Burr XII + Covariates + Spike In-sample Expected vs. Predicted

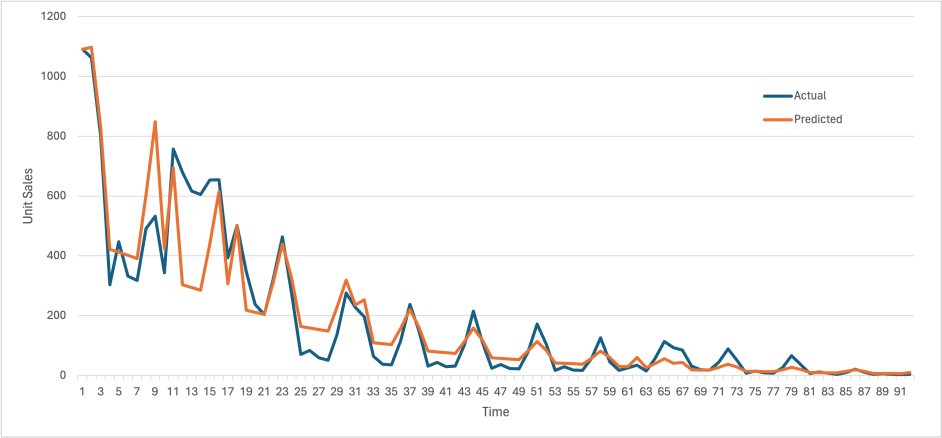

Figure 5: Weibull + Covariates + Spike In-sample Expected vs. Predicted

Parameter	Value
r	15577.48
α	532054.92
c	1.0318
λ	0.02928
p	0.3329

Coefficient	Value
β_{released}	0.3691
β_{friday}	0.4704
β_{saturday}	0.8717
β_{sunday}	0.6428
β_{holiday}	0.7803
β_{eve}	-0.4079

Metric	Performance
Log-Likelihood	-95587
BIC	191281
MAPE	0.5357

Table 3: Burr XII + Covariates + Spike Coefficients/Parameter/Performance

Parameter	Value
λ	0.0293
c	1.0318
p	0.3329

Coefficient	Value
β_{released}	0.3693
β_{friday}	0.4704
β_{saturday}	0.8717
β_{sunday}	0.6429
β_{holiday}	0.7803
β_{eve}	-0.4079

Metric	Performance
Log-Likelihood	-95587
BIC	191271
MAPE	0.5357

Table 4: Weibull + Covariates + Spike Coefficients/Parameter/Performance

The general trend of the data appears to decline exponentially while the observable moviegoers that have made an action only make up 32.948% of the population. As you see from the spiked Weibull below, disregarding the significant portion of HCNB, the results indicate that the population of buyers are extremely homogeneous and that a large portion of our population share similar propensities to buy. Strangely, this doesn't appear to be reflected in the data. For example, it appears that the general population share a similar propensity to buy more during the latter portion of the weeks. On the other hand, the December moviegoers don't appear to share the same sentiment. This provides indication of two-segment population with differing parameters despite the homogeneity telling otherwise.

Furthermore, the spike size of 33.29% reveals that, of the 32.948% which already bought, we would expect only 0.34% of the remaining 50,000 or about 170 more people that will buy tickets after March 15th. These results match real life events where Wonka (1) arrived on streaming platforms on March 8th and also (2) being nearly removed from all theatres domestically (84 theatres as of today). It's fair to assume that both would significantly reduce future individuals initiative to buy tickets. Thus, it appears that the general estimation of our model seems fairly accurate with regards to trend.

The Weibull achieved an increase of 81.43% in MAPE and 1261 decrease in BIC revealing the importance of leveraging a spike at 0. Figure 1 and the coefficients from the Weibull support the fact that Friday, Saturday, and Sunday all increase ticket sales while Saturday produces considerably more sales. Likewise, holiday's are the second biggest contributor to increasing ticket sales. As for parameters, c being slightly greater than 1 indicate little time-dependency. Intuitively, the mode claims that individual's propensity to buy doesn't quite shrink or stretch as they continue forward without having bought. Along with the insignificant c value, visual inspection of the plot reveals that the model appears to accommodate for the wave-like structure of the data while neglecting the obvious bump

around Christmas.

4 Latent Segmentation

To address the issue of apparent "Christmas fanatics" while capturing the overall patterns of the remaining population, Latent Class models provides a method to separate key differences in the population where mixture models and individual distributions fail. To reiterate, our population of "active moviegoers" appears to be classified by three different subgroups: "Christmas fanatics," "hardcore never-buyers," and the "regular moviegoers." Therefore, Latent segmentation could reveal the presence of these subgroups if they truly existed in the group of 50,000 individuals.

Since the prior models suggested a significant portion of HCNB, instead of a three-segments, we could opt for parsimony over accuracy by building a two-segment model with a spike at 0. Intuitively, due to HCNB, the three-segment model would result in one lambda extremely close to 0 to accommodate for the segment of HCNB which essentially has no propensity to buy. Therefore, the method of a spiked two-segment will save nearly 9 extra parameters. Again, these segments are built with individual Weibull distributions to account for underlying time-varying behaviors. Before discussing the details of the parameters and coefficients, the performance of the 2-segment below reveals significant improvement from the prior Weibull with the log-likelihood decreasing by 992, BIC by 1865, MAPE by 23.383%, and importantly, much better fit around Christmas and New Years.

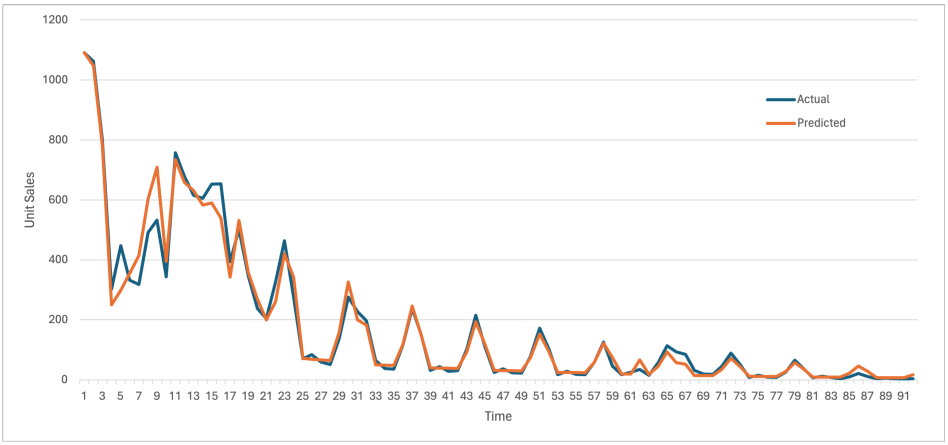

Figure 6: 2-segment Weibull Distribution + Spike In-sample Actual vs. Expected

2-Segment HCNB Metric	Performance
Log-Likelihood	-94594.7722
BIC	189405.940
MAPE	0.30187

Table 5: Spiked Two-Segment Weibull Performance

This indicates that compared to the Weibull with a spike that tries to describe the population propensity of each individual, the Latent class reveals that segmenting the population into three separate homogeneous groups is much more appropriate. For the purposes of revealing interesting subgroup specific behaviors, I created another 2-segment model with the inclusion of "scaled trend." Importantly, I omitted this feature from my other models as it doesn't change the magnitude nor direction of the other features or overall performance in any alarming manner. A closer look at the parameters and coefficients is found below:

Segment 1 Parameter	Value
λ	0.0008
c	2.7756
π	0.3644

Segment 2 Parameter	Value
λ	0.0199
c	0.9268
p	0.3398

Table 6: Segment 1 Parameters and Coefficients

Although it can be understood that λ represents the population propensity and c reveals the time-varying component, the intuition behind how these two directly relate off of purely numeric value proves far more difficult. Therefore, I have produced a plot of the general shape of each segment without the addition of covariates to provide a clear visualization of the segment's created.

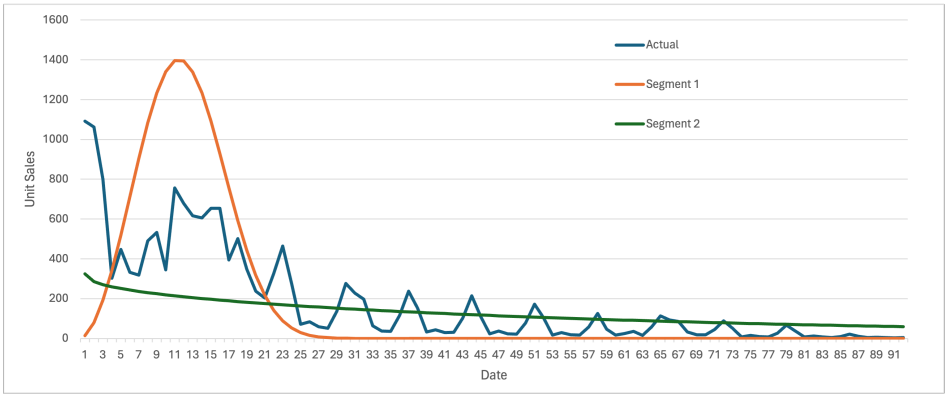

Figure 7: 2-Segment Plot Respective to True Distribution

The figure above reveals that segment 1 generally describes the "Christmas fanatics" that appear to have watched Wonka directly as a result of Christmas Break or due to the close release date. On the other hand, segment 2 appears to be the propensity of "casual moviegoers" that watch Wonka more systematically across the span of multiple months.

Segment 1 Coefficient	Value
β_{released}	5.7191
β_{friday}	-0.2311
β_{saturday}	-7.912
β_{sunday}	0.9017
β_{holiday}	-0.4646
β_{eve}	-1.1035
β_{trend}	0.1331

Table 7: Segment 1 Parameters and Coefficients

The high c value of 2.7756 for segment 1, indicates the stretching of time for people in that subgroup. Although the plot may be initially deceptive, individual's in segment 1 are almost guaranteed to watch it during the periods of Christmas to New Years. Therefore, as "Christmas fanatics" leave Christmas without watching it, their arbitrary "time" stretches to ensure they get caught in the net closest to New Years since the probability that they watch it beyond New Years is near 0. You'll notice that segment 1 reveals much different coefficients than intuition would initially lead you to believe. As seen in figure 2, notice that the weekends directly follows after the release date and precede both Christmas and New Years Eve. Since a portion of "Wonka fanatics" will likely watch the movie on release date,

the weekend following directly after release will see a subsequent drop in sales since peak viewership occurs on release. Additionally, the remaining weekends of this segment leading into Holiday eve's may indicate people will likely not attend theatres as they are preparing and spending time with their families leading into the big Holiday celebration. Therefore, I believe that the coefficients for segment 1 are much more misleading that the regular behavioral patterns in response to weekends and Holiday's.

Segment 2 Coefficient	Value
β_{released}	-121.5019
β_{friday}	0.9360
β_{saturday}	1.7423
β_{sunday}	1.3124
β_{holiday}	1.2946
β_{eve}	-148.3815
β_{trend}	-0.0050

Table 8: Segment 2 Parameters and Coefficients

Contrary to segment 1, segment 2 exhibits nearly opposite behaviors. As for the coefficient's, "Casual moviegoers" appear to watch primarily on Friday's, weekends, and Holiday's. Additionally, this population of moviegoers appears to be uninfluenced by the release date indicating that they generally may not be as interested in Wonka but if they can find the free time they will try to watch the movie. This is revealed through a similar behavior, albeit at a much larger magnitude, to not view the film on Holiday's eve. In contrast, Google trend searches exhibit practically no impact on sales for segment 2 versus the slightly positive correlation in segment 1. Perhaps there's reason to believe that the "Christmas fanatics" are more influenced by social media buzz rather than the casual moviegoer that's not up-to-date with the social media presence of the film.

5 Potential Issues

Briefly, I will discuss the decision to use a spiked 2-segment rather than a 3-segment. Notice in the performance table below, the 3-segment saw the MAPE increase by 0.1% and BIC by nearly 100, with only an negligible decrease in log-likelihood by a fraction of a point. Firstly, the results reveal that parsimony can be preferred over accuracy in certain cases, especially when the underlying population

propensity of certain groups can already be inferred. Additionally, the results reveal a large issue with increasing segments with various parameters. Although I had ran my solver multiple times at different intuitive starting points, I could not particularly guarantee that I found the absolute global maxima. This issue arises when the solver struggles to find the exact optimal single global maxima combination in the millions of combinations of β, λ, π, p, and c. Therefore, despite running the solver hundreds of times with many different starting points, it's difficult to guarantee that these Latent Class results are the global maxima.

3-Segment Metric	Performance
Log-Likelihood	-94594.74209
BIC	189503.258
MAPE	0.3029

Table 9: Three-Segment Weibull Performance

6 Conclusion and Managerial Implications

In conclusion, the results of the Weibull model with spike at 0 initially revealed that the propensity of observed Wonka moviegoers were exceptionally homogeneous. With nearly the remainder of the population being classified as hardcore never-buyers if they hadn't bought within three month of release, this emphasizes the importance of having the highest reach within the first month. Furthermore, if the Latent Class describes the true population of "active moviegoers," then the π of 0.3644 indicates that perhaps upwards of 36.44% of the profits earned may have been the result of Christmas break. From a managerial perspective, these finding presented can provide film makers an initial foundation to make more informed decisions regarding film release schedules. Additionally, the impact of increased viewing on the release date highlights the importance of marketing strategies to accumulate greater anticipation and exposure for the film.

A Appendix

The data regarding domestic theatre coverage of Wonka can be found below:

- https://www.boxofficemojo.com/release/rl2942927617/