Our data is now ready to be analyzed by applying inferential statistics techniques to explore the data. We have split up our data into five different groups focusing on the different movie franchises. These are Batman (Warner Brothers), Avengers (Marvel), X-Men (Fox), Spider-Man (Sony), and Justice League (Warner Brothers). Since each franchise on the comic book side and movie side have had their own successes and failures it is expected that each will have different results. We are analyzing how theatrical runs of movies have translated to sales of the source material it is based upon, comic books. The dataset that we have acquired was able to discover what comic books were on a given month for each of these franchises. While each franchise will sometimes have multiple titles running simultaneously, we have average the sales monthly from April 1997 and continuing until the end of year 2019.

The variables in our data we want to be focusing on is the sales of the comic books during these months and also the months that we had a movie playing. From the visualization we have done for the data, sales over time graphs shows that we have numerous peaks and valleys and no simple increasing and decreasing linear trend. The data seems to respond to some type of seasonal trend where comic books sales seem to increase at specific times and also comic book movies usually seem to be released in the summer.

Our first statistical method we are employing is finding the difference between comic book sales that occurred when a movie versus when a movie was not playing. Running a bootstrap sampling the mean 10,000 times we discover that Avengers, Justice League, and X-Men had more sales when movie were not playing while Spider-Man and Batman had more sales when movies were playing. What’s interesting from these results is that not only are Batman and Spider-Man the only franchises that had more sales when movie were playing but they also had the lowest difference in means for movie vs no movie. Batman had 4137 and Spider-Man had 1090 difference in mean while the other franchises had differences of around -20000.

Next for our statistics analysis we will be conducting a hypothesis test so that we can either accept or reject based on our findings. There will be two kinds of hypotheses: the null and the alternate. The null hypothesis is the control and is assumed to be true usually with some sort of equal sign while the alternate is the opposite of the null and what we usually want to be proven true. For our purposes our null hypothesis will be that there is no difference in comic book sales when a movie was playing versus when it was not playing. The alternate hypothesis will be the opposite of this statement in that that is a difference in comic book sales when a movie playing versus having no movie playing.

The way that we test this hypothesis is with bootstrapping and confidence intervals that will check if what we are looking for is consistent with what we observe in the sampling distribution. We will take the difference in means we previously calculated and then select values at the chosen percentile for the confidence interval. The chosen percentile is called alpha and we have chosen a confidence interval of 95%. After we have calculated our confidence interval from our bootstrapped difference of means, we make a histogram plot showing the distribution and drawing vertical lines showing where the intervals lie on. What we observe is that even though we know that the difference in mean for Spider-Man was over 1000 when looking at the plot we see that that a difference in zero is in the middle of the histogram. Even Batman which had a slightly higher difference of 4000 does not zero to in the middle of the histogram.

While we have two hypothesis that we are testing for, so far we do not have a way to test whether or not they are true or false. For that we need a probability value which is more commonly known as the p-value. If the null hypothesis is true then the p-value is the probability of obtaining your sample data. At a certain point we can make the determination that the null hypothesis is false since the p-value is the probability of our data occurring. This point is called the significance level and is represented by alpha which we already defined with the confidence interval. It is another way of saying how willing we are to committing errors and in our case we are willing to have a 5% rate. When our p-value goes below the value of 0.05 we can reject our null hypothesis.

To obtain our p-value we use the SciPy stats T-test comparing the movie and no-movie means. This is a two sided test for the null hypothesis for two independent samples which we calculate with different means and variances which we discovered previously. We are able to reject the null hypothesis for Avengers, Justice League, and X-Men since their p-value is extremely small. However, we are unable to reject the hypothesis for Batman and Spider-Man since their p-values are so large.

Our results show that depending on the franchise we can have either a difference in means or no difference when looking at the separate groups of comic book sales for when a movie playing versus no movie playing. This means that when we are modeling our data and creating predictions for them, we should be getting different results for each franchise.