Final Project Steps 1-3

Taylor Woodington

5/18/2024

Introduction

Worldwide, there are many different theme parks and water parks that people often use as summer or holiday vacation. Among these are some of the more popular parks such as Disneyland/World and Universal. These two are known competitors with parks and reside close to one another in California and Florida with park attendance that may fluctuate due to the other parks or attractions around them. The problem I plan to address is which is receiving more attendance and what do their review scores rate these parks at. Some dedicated park goers and individuals of each company that runs each park may be interested to see what the data shows and how their competitors stack up against them in attendance and reviews. This is competition, and if it shows their competitors are doing better, the companies may want to investigate why. They could look back at when some attractions were introduced or updated and how the attendance was during that time. Data science can be used to see if those competitors are doing better, the parks' attendance can be compared to one another over the years as well as the parks reviews. The higher the attendance, the more tickets purchased and the more opportunities for purchases of goods, the better reviews, the more possibilities of guests returning or new guests coming in.

Research Questions

1. Which Park has had the most increasing attendance over the years listed in the dataset?
2. Do either of the parks show a period of attendance decreasing over more than two years?
3. Which Park has had the longest attendance increase? (years in a row)
4. Is there a strong correlation between the attendance of Universal Studios and Disneyland California?
5. Is there a relationship between ratings and years of the ratings?
6. Which years seem to show to have the most significant impact?

Approach

First, I plan to clean up the data. I will do this by combining the Disneyland attendance table along with the Universal Studios attendance table to have them next to one another to do a covariance and correlation matrix to see if there is a possible relationship or pattern between the two parks over the years. For the reviews data sets, I will clip out missing reviews and narrow the review years between 2015 and 2019 that way there's data for both parks. Once the data is cleaned up, I plan to test the data by using a histogram for each of the park's attendance to see where it is skewed, or the data points are clustered the most. Afterwards I plan to use covariance and correlation matrices to see if there is a relationship between the years and the

parks attendance. Then, I plan to do a simple regression between the ratings and years of both data sets to see if the ratings depend on the year. These tests may be able to show if there is a relationship between the two parks and their attendance or if there is a relationship between the years and the reviews.

How your approach addresses (fully or partially) the problem

My approach will partially solve the problem. While the approach I am taking will not show exact reasons the park attendance changes or why the reviews may be higher or lower in certain years, it may spark some thought or investigation into those specific years. By knowing what relationships may exist or what years may be of importance, others can take that data and research what attractions may be added to the parks or updated and rereleased to the public. By learning those things, the separate businesses may be able to strategize what is important to guests, what years were successful, and what those years consisted of.

Data (minimum of 3 datasets)

Review datasets seemed to be used out of interest and from visitors posting on Trip Advisor, whereas the attendance datasets were taken from the AECOM Theme Index of years 2006 to 2022.

1. Disneyland Reviews

-Columns include an individual who wrote the review, reviews, date of the review being input, and a comment

-Reviews range from 1 (unsatisfied) to 5 (satisfied)

-About 6% of values are missing the year the review was put in -About 42,632 entries of reviews ranging between the years of 2011 and 2019

-Found on Kaggle, data collected from Trip Advisor Dataset put together by Arush Chillar

`https://www.kaggle.com/datasets/arushchillar/disneyland-reviews?resource=download`

2. Universal Studios Reviews

-Columns include an individual who wrote the review, reviews, date of the review being input, and a comment

-Reviews range from 1 (unsatisfied) to 5 (satisfied)

-About 50,847 entries of reviews ranging between the years of 2010 and 2021

-Found on Kaggle, data collected from Trip Advisor Dataset put together by Dwi Gustin Nurdialit

`https://www.kaggle.com/datasets/dwiknrd/reviewuniversalstudio`

3. Disneyland Claifornia Historical Attendance Data

-Columns include the year and attendance

-Shows Park attendance between the years of 2006-2022

-Found on Queue Times, data collected from AECOM Theme Index between the years 2006 to 2022

`Queue Times: https://queue-times.com/en-US/parks/16/attendances`

`AECOM Theme Index: https://aecom.com/wp-content/uploads/documents/reports/AECOM-Theme-Index-2022.pdf`

4. Universal Studios Orlando Historical Attendance Data

-Columns include year and attendance

-Shows Park attendance between the years of 2006-2022

-Found on Queue Times, data collected from AECOM Theme Index between the years 2006 to 2022

`Queue Times: https://queue-times.com/parks/65/attendances`

`AECOM Theme Index: https://aecom.com/wp-content/uploads/documents/reports/AECOM-Theme-Index-2022.pdf`

Required Packages:

Main required packages include ggplot2 for making plots, readxl for reading the data sets as excel sheets, and dplyr to make data manipulation easier.

Plots and Table Needs

A table including the years and park attendance will be used along with the other datasets which include reviews and what years those reviews were put in. Plots and tables to help with illustration will include multiple histograms as discussed before, covariance and correlation matrices, as well as simple regression models with residuals plotted to see if the regression model is appropriate or helpful.

Questions for Future Steps

As of right now I believe that I would need to learn a little bit more about easier ways to compare the data or more ways to combine the datasets and look for different relationships. I feel as if there is another comparison I can be making that I am unaware of. Possibly comparing a dataset to another without it being combined and doing some kind of multilinear regression model that may incorporate both review datasets with the attendance dataset, making the year the explanatory variable with the attendance and ratings as the dependent variable.

How did you import and clean your data?

First, I pulled up all the data sets that I planned to use. I started with the Disneyland California and Universal Studios Orlando historical data attendance sets. Due to this part of the data being relatively smaller than the rest, I copy and pasted the years I planned to look at (2011-2022) into a new blank sheet on a blank excel workbook. This workbook is labeled as "park_attendance"

Then I downloaded each of the park reviews data sets from both Universal Studios and Disneyland. First, I cleaned up the Universal Studios data by taking out the review title column and comments column, this is data that will not be used, seeing it will be a lot of writing that can be categorized by the satisfaction rating. I am looking to use the numerical data of the actual satisfaction rating out of 5.

After cleaning up the Universal Studios data, I did something similar with the Disneyland Reviews data set. I took out the columns about the individual reviewers location and the written review column (this is not the data being analyzed, as stated above). Again, I am looking to use the numerical data of satisfaction out of 5 to categorize the reviews.

From there I filtered the data sets using code as shown and explained below: As shown, the pipe operator was used to make the code more efficient rather than each pipe function being multiple separate functions.

```
# Load necessary libraries
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.1.1
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.1
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 4.1.1
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(zoo)
```

```
## Warning: package 'zoo' was built under R version 4.1.3
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
# Import Universal Reviews Data
universalreviews <- read_excel("C:/Users/Shaun/Downloads/UniversalReviews.xlsx")

# Import Disneyland Reviews Data
disneylandreviews <- read_excel("C:/Users/Shaun/Downloads/DisneylandReviews.xlsx")

# Import Park Attendance Data
park_attendance <- read_excel("C:/Users/Shaun/Downloads/park_attendance.xlsx")

# Filter Universal Review Data
# Trim data to only include reviews about Universal Studios Florida
# Separate the year from the date column and filter to show years 2011-2019 (slice and dice data)
# Remove the original date column and the column which includes the reviewers name
filtered_universalreviews <- universalreviews %>%
  filter(trimws(branch) == "Universal Studios Florida") %>%
  mutate(year = format(as.Date(date), "%Y")) %>%
  filter(between(as.numeric(year), 2011, 2019)) %>%
  select(-date, -reviewer)

# Filter Disneyland Reviews Data
# Trim data to only include reviews about Disneyland California
# Format the Year_Month column to be a date instead of text
# From the date column pull out the year
# Organize the years to be between 2011 and 2019 (slice and dice data)
# Remove the columns that are Year_Month and Date, as well as the Reviewer_ID
filtered_disneylandreviews <- disneylandreviews %>%
  filter(trimws(branch) == "Disneyland_California") %>%
  mutate(Date = as.Date(as.yearmon(Year_Month, "%Y-%m")),
         year = format(Date, "%Y")) %>%
  filter(between(as.numeric(year), 2011, 2019)) %>%
  select(-Year_Month, -Date, -Review_ID)

# Combined the two review data sets
combined_parkreviews <- rbind(filtered_universalreviews, filtered_disneylandreviews)
```

What does the final data set look like?

Below I show a condensed version of both review data sets for the separate parks as well as the final clean data set of them combined. In addition, I show the park attendance data set.

```
# Show what slice and diced Universal reviews look like
head(filtered_universalreviews, 10)
```

```
## # A tibble: 10 x 3
##    rating branch                    year
##     <dbl> <chr>                     <chr>
##  1      5 Universal Studios Florida 2019
##  2      5 Universal Studios Florida 2019
##  3      4 Universal Studios Florida 2019
##  4      5 Universal Studios Florida 2019
##  5      5 Universal Studios Florida 2019
```

```
##  6      4 Universal Studios Florida 2019
##  7      5 Universal Studios Florida 2019
##  8      5 Universal Studios Florida 2019
##  9      5 Universal Studios Florida 2019
## 10      5 Universal Studios Florida 2019
```

```
# Show what the slice and diced Disneyland reviews look like
head(filtered_disneylandreviews, 10)
```

```
## # A tibble: 10 x 3
##    rating branch               year
##     <dbl> <chr>                <chr>
##  1      5 Disneyland_California 2019
##  2      5 Disneyland_California 2019
##  3      4 Disneyland_California 2019
##  4      5 Disneyland_California 2019
##  5      5 Disneyland_California 2019
##  6      5 Disneyland_California 2019
##  7      5 Disneyland_California 2019
##  8      3 Disneyland_California 2019
##  9      5 Disneyland_California 2018
## 10      5 Disneyland_California 2019
```

```
# Show Park Attendance Data set
head(park_attendance, 12)
```

```
## # A tibble: 12 x 3
##    Years `Disneyland California Attendance` `Universal Studios Orlando Attendan~
##    <dbl> <chr>                              <chr>
##  1  2022 16881000                           10,750,000
##  2  2021 8,573,000                          8,987,000
##  3  2020 3,674,000                          4,096,000
##  4  2019 18,666,000                         10,922,000
##  5  2018 18,666,000                         10,708,000
##  6  2017 18,300,000                         10,198,000
##  7  2016 17,943,000                         9,998,000
##  8  2015 18,278,000                         9,585,000
##  9  2014 16,769,000                         8,263,000
## 10  2013 16,202,000                         7,062,000
## 11  2012 15,963,000                         6,195,000
## 12  2011 16,140,000                         6,044,000
```

```
# Provide the clean data set of the reviews for both parks including the year and various filters
head(combined_parkreviews, 22)
```

```
## # A tibble: 22 x 3
##    rating branch                   year
##     <dbl> <chr>                    <chr>
##  1      5 Universal Studios Florida 2019
##  2      5 Universal Studios Florida 2019
##  3      4 Universal Studios Florida 2019
##  4      5 Universal Studios Florida 2019
```

```
##  5      5 Universal Studios Florida 2019
##  6      4 Universal Studios Florida 2019
##  7      5 Universal Studios Florida 2019
##  8      5 Universal Studios Florida 2019
##  9      5 Universal Studios Florida 2019
## 10      5 Universal Studios Florida 2019
## # ... with 12 more rows
```

With cleaning the data above, I had to do more work and research on pipe functions in order to make sure that the function did each step accurately and in the right order.

What information is not self-evident?

When looking at the data there are thousands and thousands of observations which can make information hard to see by just a glance. I plan to uncover relationships between the years and ratings to see if any years have a stronger significance when compared to other parks. I plan to see their significance through linear regression models.I plan to look at a linear regression model of the parks separately with years and rating as well as them together overall to see if there is a difference in relationships.

What are different ways you could look at this data?

When looking at the different data I think starting off with simple plots and histograms could be very useful in answering the first couple of questions I have that could be based off of a well labeled visual. For questions such as "Which park has had the most increasing attendance over the years listed in the data set?" and "Do either of the parks show a period of attendance decreasing over more than two years?" those visuals can show these things, and then two simple linear regressions may show if there is a relationship between the year and attendance. When looking at the ratings of data something similar can be done as well as the use of tables to show summary statistics or a summary of linear regression to show significant years.The data can be looked at through some visuals, plots, and small tables.

How do you plan to slice and dice the data?

As shown above, I have already begun to slice and dice the data by condensing which years I will be using for data analysis. In addition, I have removed columns which I feel are not significant in being shown in the data frames that could be a distraction from the data being analyzed. Moving forward, I plan to use a combined data frame of both the Universal Reviews and Disneyland Reviews to see if there is a relationship or pattern between the two. In addition, I plan to look at the three smaller data sets individually to make histograms for each of the years to have a better visual of the data in one place without a large clump of numbers. This way I can get a summary of the reviews together as well as separate, in addition to a summary of park attendance over the years.

How could you summarize your data to answer key questions?

I could summarize my data to answer key questions by not only using the summary tool for each of the data frames, but also by using visuals and plots to summarize the data. When there is a plot or model I can explain what the visual shows to summarize the importance of increasing or decreasing attendance and use the tables to explain which years may be of interest for future studies. In addition, the use of linear regression models will create a summary showing whether different variables have a relationship with one another.I plan to use simple plots and tables as well as linear regression models to summarize key questions.

What types of plots and tables will help you illustrate the findings to your questions?

Below I have shown the types of plots and tables to help illustrate the findings to questions. Each have comments in order to understand what they are being used for or the relationships being analyzed in the situation.

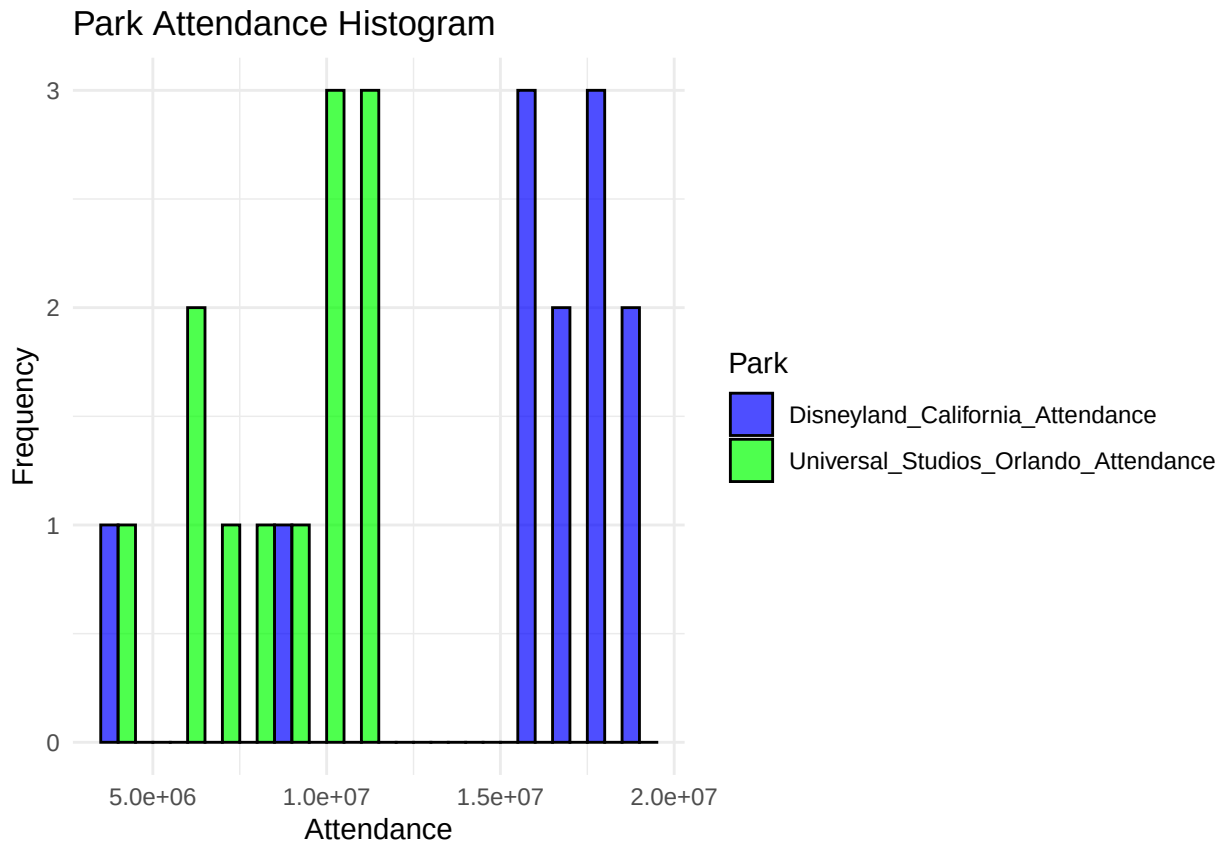
```
# Load necessary libraries
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.1.1
```

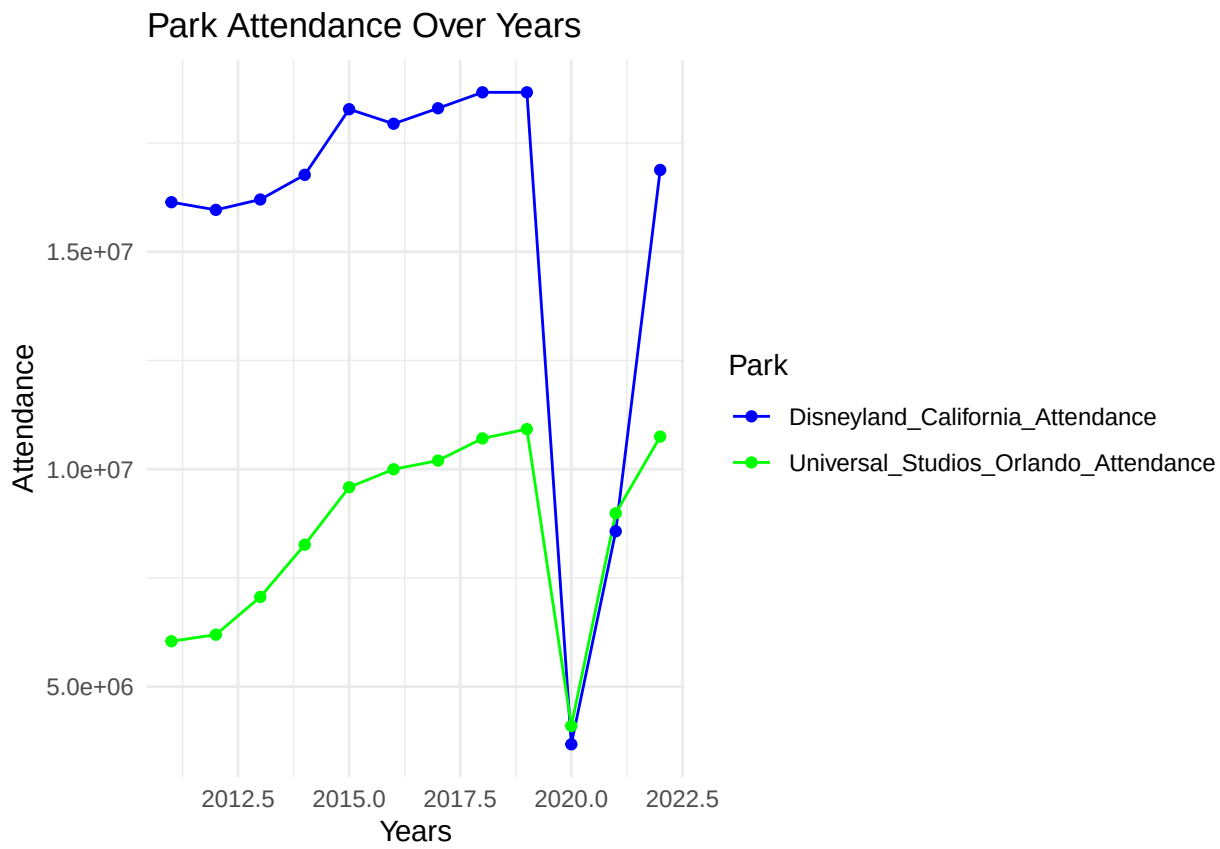
```
# Create an attendance data frame to better organize the park attendance data with variables that can b
attendancedata <- data.frame(
  Years = c(2022, 2021, 2020, 2019, 2018, 2017, 2016, 2015, 2014, 2013, 2012, 2011),
  Disneyland_California_Attendance = c(16881000, 8573000, 3674000, 18666000, 18666000, 18300000, 179430
  Universal_Studios_Orlando_Attendance = c(10750000, 8987000, 4096000, 10922000, 10708000, 10198000, 995
)

# Reshape the data from wide to long format
attendancedata_long <- tidyr::pivot_longer(attendancedata, -Years, names_to = "Park", values_to = "Atte

# Plot the histogram of attendance between the two parks
ggplot(attendancedata_long, aes(x = Attendance, fill = Park)) +
  geom_histogram(binwidth = 1000000, position = "dodge", color = "black", alpha = 0.7) +
  labs(x = "Attendance", y = "Frequency", title = "Park Attendance Histogram") +
  scale_fill_manual(values = c("blue", "green")) +
  theme_minimal()
```




```
# Plot the attendance of the parks between the years 2011 and 2022
# This will be a very simple but useful visual to understand the pattern over time in attendance for ea
ggplot(attendancedata_long, aes(x = Years, y = Attendance, color = Park, group = Park)) +
  geom_line() +
  geom_point() +
  labs(x = "Years", y = "Attendance", title = "Park Attendance Over Years") +
  scale_color_manual(values = c("blue", "green")) +
  theme_minimal()
```



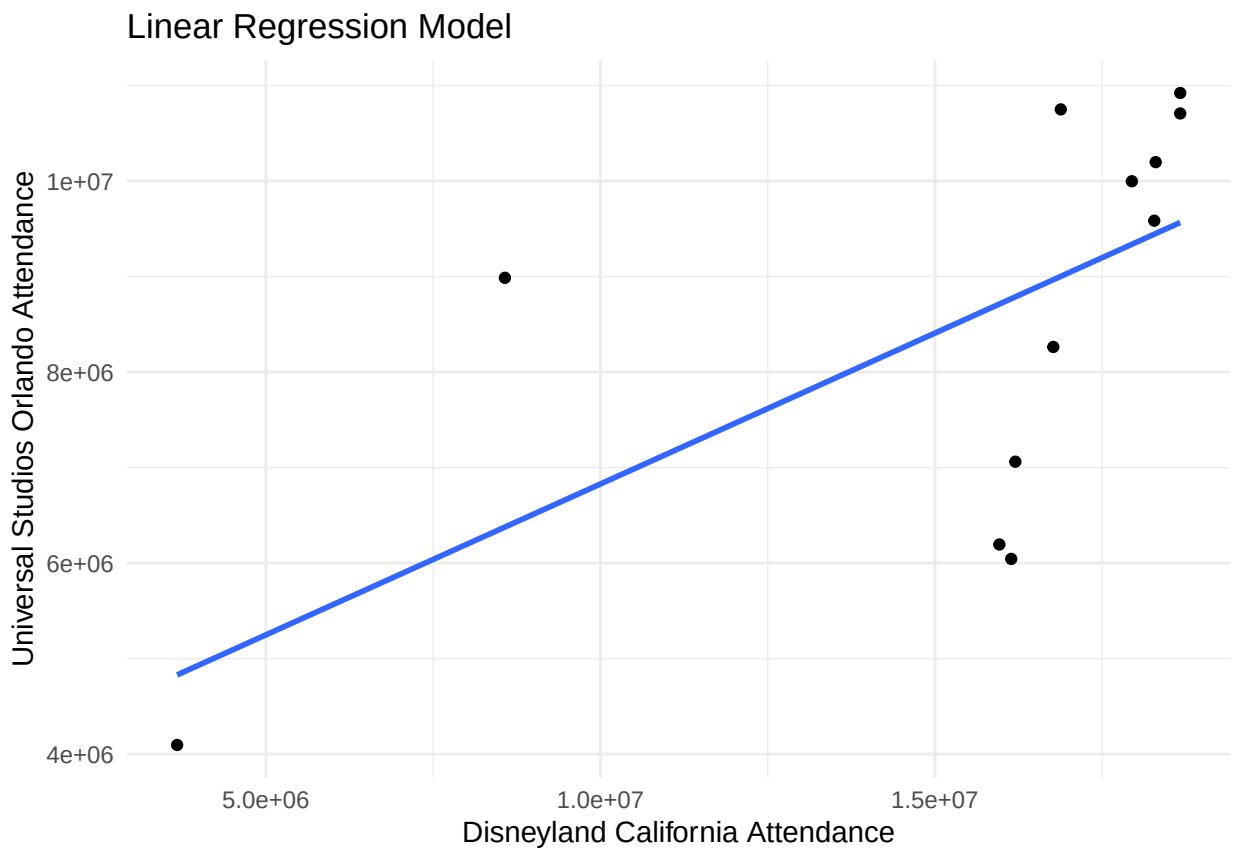

```
# A linear regression model of the attendance for both parks
# Fit a linear regression model
lm_attendance_model <- lm(Universal_Studios_Orlando_Attendance ~ Disneyland_California_Attendance, data

# Print the summary
summary(lm_attendance_model)
```

```
##
## Call:
## lm(formula = Universal_Studios_Orlando_Attendance ~ Disneyland_California_Attendance,
##     data = attendancedata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2724101  -981406   400794  1195277  2609788
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     3.668e+06  1.879e+06   1.952   0.0794 .
## Disneyland_California_Attendance 3.160e-01  1.165e-01   2.711   0.0219 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1783000 on 10 degrees of freedom
## Multiple R-squared:  0.4236, Adjusted R-squared:  0.366
## F-statistic:  7.35 on 1 and 10 DF,  p-value: 0.02189
```

```r
# Create a scatter plot of the data points
# Creating a visual of model
ggplot(attendancedata, aes(x = Disneyland_California_Attendance, y = Universal_Studios_Orlando_Attendanc
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Disneyland California Attendance", y = "Universal Studios Orlando Attendance", title = "Lin
  theme_minimal()
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```r
# Fit a linear regression model for both parks ratings
lm_parkreviews_model <- lm(rating ~ year, data = combined_parkreviews)

# Print the summary for analysis
summary(lm_parkreviews_model)
```

```
##
## Call:
## lm(formula = rating ~ year, data = combined_parkreviews)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4254 -0.3668  0.5746  0.6712  0.8315
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.272244   0.026360 162.074  < 2e-16 ***
## year2012     0.094542   0.031740   2.979 0.002896 **
## year2013     0.047417   0.030572   1.551 0.120910
## year2014     0.056592   0.029518   1.917 0.055217 .
## year2015     0.153181   0.028524   5.370  7.9e-08 ***
## year2016     0.105473   0.028621   3.685 0.000229 ***
## year2017     0.089006   0.029313   3.036 0.002396 **
## year2018     0.003345   0.029791   0.112 0.910606
## year2019    -0.103776   0.032555  -3.188 0.001435 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.023 on 47031 degrees of freedom
## Multiple R-squared:  0.003988,   Adjusted R-squared:  0.003819
## F-statistic: 23.54 on 8 and 47031 DF,  p-value: < 2.2e-16
```

```
# Fit a linear regression model for the Disneyland Califronia ratings only
lm_disneylandparkreviews_model <- lm(rating ~ year, data = filtered_disneylandreviews)

# Print the summary for analysis
summary(lm_disneylandparkreviews_model)
```

```
##
## Call:
## lm(formula = rating ~ year, data = filtered_disneylandreviews)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5429 -0.4075  0.5271  0.5925  0.7568
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.48793    0.03123 143.726  < 2e-16 ***
## year2012     0.05492    0.03711   1.480  0.13890
## year2013    -0.01499    0.03707  -0.404  0.68585
## year2014    -0.08044    0.03642  -2.209  0.02722 *
## year2015    -0.06126    0.03552  -1.724  0.08465 .
## year2016    -0.09686    0.03615  -2.679  0.00738 **
## year2017    -0.17446    0.03775  -4.621 3.84e-06 ***
## year2018    -0.24476    0.03993  -6.130 8.95e-10 ***
## year2019    -0.13177    0.06448  -2.044  0.04100 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.964 on 18121 degrees of freedom
## Multiple R-squared:  0.00704,    Adjusted R-squared:  0.006602
## F-statistic: 16.06 on 8 and 18121 DF,  p-value: < 2.2e-16
```

```
# Fit a linear regression modelof the UniversalStudios Florida ratings only
lm_universalandparkreviews_model <- lm(rating ~ year, data = filtered_universalreviews)
```

```r
# Print the summary for analysis
summary(lm_universalandparkreviews_model)
```

```
##
## Call:
## lm(formula = rating ~ year, data = filtered_universalreviews)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4247 -0.3842  0.5753  0.7120  1.0995
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.90054    0.04465  87.366  < 2e-16 ***
## year2012     0.07441    0.05527   1.346    0.178
## year2013     0.24386    0.05035   4.844 1.28e-06 ***
## year2014     0.36505    0.04826   7.565 4.00e-14 ***
## year2015     0.52416    0.04681  11.198  < 2e-16 ***
## year2016     0.47052    0.04679  10.056  < 2e-16 ***
## year2017     0.48364    0.04743  10.197  < 2e-16 ***
## year2018     0.38742    0.04768   8.125 4.65e-16 ***
## year2019     0.24664    0.04921   5.012 5.41e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.05 on 28901 degrees of freedom
## Multiple R-squared:  0.01415,    Adjusted R-squared:  0.01388
## F-statistic: 51.85 on 8 and 28901 DF,  p-value: < 2.2e-16
```

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

I plan to incorporate the machine learning technique of linear regression into my research questions the most. This is because it can be used to model the relationship between attendance with the years. As well as using it to model relationships between the satisfactory ratings and the years or attendance of those years. The regression can show if different years have different significance to the ratings in order to provide interest years for future interpreters to study.

What questions do you have now, that will lead to further analysis or additional steps?

Currently, I am questioning if I should add the overall year park attendance into the data set of combined park reviews. I could add it as an additional column in order to test the relationship between the satisfaction scores and the attendance that year in a linear regression model. I am just unsure how to transfer and match the year and park to the proper row or if it will make the data too confusing to analyze. This question is leading to additional analysis and possible additional steps.

Overall Narrative

Introduction, Problem Statement and How it was Addressed

There are many different parks throughout the country that people travel to for a good vacation with friends or family. Some of the busiest and most well known parks are The Disneyland and Universal Studios parks. I believe these parks are often considered as competition for each other, so I wanted to compare the two. There are multiple branches to each park, and I decided to choose the one from each that had data for the United States since that is where I live. This lead me to choosing Disneyland California and Universal Studios Florida. It was thought provoking to think of which park is more successful or popular. The problem statement I planned to address was which park between the two is receiving more attendance and higher ratings through scores from guests who have attended the parks.

For each of the parks, I had a data set including multiple variables which I narrowed down to show individual reviews and which year they were written for each park, as well as a separate dataset that included the attendance for each year and each park between 2012 and 2022. In order to address both parts of the problem statement I did analysis on the attendance data set first. I used a histogram and plot with a linear regression model to show the relationship between attendance and year for Disneyland California and Universal Studios Florida. In order to look at the relationship between the ratings over the years for each of the parks I did separate linear regression models for each of the parks to show how the ratings changed over the years for each park. Another model I would recommend for the possible future is a correlation or regression model that pairs the reviews against the attendance for the year that review was written. In addition, if there was access to more data such as cost it would be interesting to see the relationship between the reviews and the cost of the ticket for each park.

Analysis and Implications

By looking at the simple plot and linear regression model, I can see that as the years increase, attendance in both parks increase, with the exception of 2020 when both parks shut down due to the pandemic. However, prior to the shut down it shows that while both parks were successfully increasing their attendance, Disneyland had a much higher attendance than Universal did. After the shutdown I notice both parks attendance jumps up, with Disneylands attendance higher than Universals. However, Disneylands attendance does not increase from the number that it was before the pandemic or reach that same number, wheras Universal Studios does reach the same number as before the pandemic. The regression model shows as years increase and attendance at Universal increases, the attendance at Disneyland increases as well. This means both parks may be successful at keeping interest. While Disneyland's attendance is a higher attendance than Universal, when compared to itself in previous years before the pandemic, it is less. This implies that Disneyland may not be doing as well with attendance as it seems when only compared to Universal Studios.

After looking at the attendance, I made two separate linear regression models for Disneyland California and Universal Studios Florida to look at the relationship between years and ratings. When looking at the results from the Disneyland regression model I notice that after the year 2014, the coefficients continue to decline as well as the p-values, p-values of the year 2014 and 2016 to 2019 are all significant p-values. Overall this regression model, suggests that Disneyland ratings have been decreasing over the years, especially after 2014. However, after looking at the R-squared value, I can see it is really low which may suggest that the year is not enough to explain the change in ratings. Alternately, Universal Studios Florida shows a different trend in their linear regression model. The coefficients for years 2013 to 2019 are all statistically significant and indicate a trend where the ratings increase over the years. However, this R-squared value is low here as well and suggests that the year alone may not be enough to explain the trend being seen.

Overall, the data has shown that Disneyland, while keeping a increasing attendance (which again, is less than before the pandemic), is declining in reviews, wheras Universal Studios is keeping an increasing attendance as well as increasing reviews. This implicates that Universal Studios may be considered to be doing better than Disneyland is doing. This is important for the target audience, individuals running the parks itself,

because they care about the competition and reviews. The parks are open for the guests, and if reviews show that they are not enjoying their time as much as they used to this may affect the attendance in the future even though it may not have yet. There has been some kind of decline in attendance when taking out the year of the pandemic when analyzing Disneyland, so the consistency of more negative reviews over years could change attendance if needs of the guests are not met. The target audience may want to look into what guests are not in favor of to consider changing it.In addition, if the park is doing well like Universal Studios Florida, then they may want to assess what guests like and continue to do it and improve upon the rest.

Limitations and Concluding Remarks

While the analysis done in this project does spark interesting conversation and thoughts, I do not believe the analysis is done. The data that I had to work with was a little limiting and could be better if there were more variables involved such as the average cost of tickets during each year or which years had attraction releases to the public. In addition, it is difficult to compare these two parks when they are across a country from each other. While they are in the same country, they are far away from one another and different populations have different types of access to them. To improve an analysis like this, it would be good to have data on Walt Disney World in Florida. This way, both parks would be located in a very similar area with similar levels of access to the public. I could not find data such as the ones I have described, so compiling this data would be necessary in order to do those types of analysis. With the data that I did have, I believe another model that someone could do to test different relationships could be the correlation relationship between the rating and the years attendance that the rating was written. This could show if ratings are possibly affected by the attendance or vice versa. I believe that future analysis would be beneficial for this type of study and could show the audience different years of interest to dive deeper into. If there are years with a large decline in population or reviews, they may want to know why in order to fix the issue causing a decline. From the current relationships shown, I believe Disneyland California should focus on what was around before the pandemic and what is not there now that could be changing the way that those attending are viewing the park and its experience. In addition, Universal Studios Florida should look at what they have done differently to continue the same trend of improvements.