**ANALYZING THE DISNEYLAND REVIEWS**

**ABOUT THE DATA**

My dataset comprises information on user reviews of three different Disney Park branches, through the years 2010 to 2019. The three different Disney Park locations are California, Hongkong, and Paris. I have collected this dataset from Kaggle ([LINK](https://www.kaggle.com/arushchillar/disneyland-reviews)).

This dataset contains a total of 42,657 records and 6 attributes.

**DATA DICTIONARY**

|  |  |  |
| --- | --- | --- |
| **ATTRIBUTES** | **DESCRIPTION** | **ATTRIBUTE TYPE** |
| Review\_ID | Unique reviewer ID | Integer |
| Rating | Rating is given by the reviewer to that particular Disneyland branch. (1 – Unsatisfied to 5 – Satisfied). | Integer |
| Year | Year when the reviewer visited the park. | Integer |
| Month | The month when the reviewer visited the park. | Integer |
| Reviewer\_Location | Country of origin of the reviewer. | Text |
| Review\_Text | Reviews made by the visitors. | Text |
| Disneyland\_Branch | The location of Disneyland. | Text |

**GOAL OF THE ANALYSIS**

The main aim of this analysis is to perform text analysis and draw useful conclusions. I want to find out the behaviors, characteristics, and patterns in the reviews presented by the visitors throughout the different Disneyland locations.

This analysis will help the employees at the different Disneyland locations in understanding, what people think about that particular Disneyland location. They could analyze what most people like and maybe try to further improve them even more and also analyze what people don’t like and could intend to perform actions in such a way that those dislikes never happen again.

**DATA MINING TECHNIQUES USED**

Throughout this analysis I have used a total of 3 data mining techniques, those are:

1. Correlation analysis
2. Association analysis and
3. K-Means clustering analysis.

**CORRELATION ANALYSIS**

Process Diagram

Diagram

Description automatically generated with medium confidence

Correlation Matrix

Graphical user interface, application, Teams

Description automatically generated

Table

Description automatically generated

Since correlation talks about the strength or the similarity between the attributes, we could probably say that the similarity lies within different rides, themes, different lands (tomorrow land, fantasy land, frontier land, adventure land, etc.), shows, shops, or maybe even hotels and restaurants throughout the different Disneyland locations.

Disneyland\_california and Disneyland\_HongKong are correlated at -0.497. There exists a moderate relationship between those two attributes. Since they are negatively correlated, we could probably assume that either or all of the rides, theme parks, shows, shops, hotels, and restaurants are better at one Disneyland location when compared to others.

The correlation between Disneyland\_paris and Disneyland\_HongKong is -0.371, which implies that there exists a very weak similarity between these attributes. And Disneyland\_Paris and Disneyland\_California are correlated at -0.622, i.e., there exists a strong correlation between these two attributes. Since there is a negative sign, we could probably say that one of these two Disneyland locations is better when compared with the other location.

**ASSOCIATION ANALYSIS**

Association rules were formed considering all different Disneyland locations.

Process Diagram

Diagram

Description automatically generated

There were a lot of associations rules being formed, so I've set the minimum confidence to 0.4

Text

Description automatically generated

Diagram

Description automatically generated

**FINDING THE ASSOCIATION RULES BASED ON DIFFERENT LOCATION**

**Diagram

Description automatically generated**

**Graphical user interface, application

Description automatically generated**

For selecting the different locations, I have imputed the filter example operator (represented as Filter Examples (location selector)) and changed the location each time.

**Association rules for Disneyland\_California**

With minimum confidence of 0.2, there were a total of 2398 association rules being formed. So, to get the top rules, I've increased the confidence to 0.8, then a total of only 28 rules were formed.

A screenshot of a computer

Description automatically generated with medium confidence

I, even more, increase the minimum confidence to produce an interpretable graph of the association rules. The below graph is formed by increasing the confidence to 0.9.

Diagram

Description automatically generated

**Association rules for Disneyland\_HongKong**

I have used the minimum confidence of 0.8 and the following association rules were generated. I.e., a total of 72 rules have been formed.

A black and white document

Description automatically generated with low confidence

A black and white document

Description automatically generated with low confidence

I have even more increased the minimum confidence value to get a better interpretable graph. I.e., I have increased the minimum confidence to 1.

Diagram

Description automatically generated

**Association rules for Disneyland\_Paris**

I have used the minimum confidence of 0.8 and a total of 831 association rules were formed. So, I've increased the minimum confidence to 0.9 to get the top association rules. Then a total of 10 association rules were formed.

Text, letter

Description automatically generated

Diagram

Description automatically generated

As you can observe from the association rules, all the association rules created with a minimum confidence of 0.9 are talking about the thunder mountain ride in Paris.

**K-MEANS CLUSTERING**

Considering all the different Disneyland locations.

I wanted to split the data into a total of 4 clusters.

Process Diagram:

Diagram

Description automatically generated

Diagram

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Chart

Description automatically generated with medium confidence

Table

Description automatically generated with medium confidence

Based on the attributes, I’ve named Cluster\_ 0 “Putting Words to Emotions”, after arrainging them in descending order.

Table

Description automatically generated with medium confidence

I’ve named Cluster\_1 “ Expressing the Experiences “ based on its attributes, after arrainging them in descending order.

A screenshot of a computer

Description automatically generated with low confidence

I’ve named Cluster\_2 “ Defining the Distinctions “ based on its attributes, after arrainging them in descending order.

A screenshot of a computer

Description automatically generated with medium confidence

I’ve named Cluster\_3 “ Lines and Rides “ based on its attributes, after arrainging them in descending order.

**CONCLUSION:**

1. Performing correlation analysis showed us that, there exists a similarity between the different Disneyland locations.
2. By performing the association analysis, the following are the most spoken about in reviews throughout different Disneyland locations:

* The lines being too long before the rides.
* Even though the lines were too long, they were moving at a good pace.
* The food is expensive.
* The “space mountain ride” because most people enjoy it.

1. From analyzing the clusters, it is clear that

* Cluster 1 talks about the different emotions the visitors felt after experiencing Disneyland, so I’ve named that cluster “PUTTING WORDS TO THE EMOTIONS”.
* Cluster 2 talks about the experiences and the wonderful time the visitors had, so I’ve named that cluster “EXPRESSING THE EXPERIENCES”.
* Cluster 3 mainly describes the differences between Disneyland and Disneyworld, so I’ve named that cluster “DEFINING THE DISTINCTIONS “.
* And Cluster 4 talks about the huge lines and different rides, so I’ve named that cluster “LINES AND RIDES”.

**EXPERIENCES**

* RapidMiner was simple to use and to perform text analysis.
* The only difficulty that I faced was, RapidMiner operators were taking a long time to run, with the huge amount of data.
* Additionally, being able to analyze 42,657 reviews quickly was amazing.