Srisai Kumar Resu (11744462)

INFO 5810 Summer 2024

Term Project

Airline Reviews Analysis

This project focuses on analyzing an extensive dataset which contains actual airline reviews data. This analysis helps to understand the customer needs who travel with different airlines which can be used to improve the airline services and their customer experiences.

Introduction

- Air travel is a widely used mode of transportation globally, preferred for its speed and efficiency over long distances.
- Understanding customer needs and experiences is vital for improving airline services and increasing revenue.
- By analysing customer reviews, airlines can identify areas of improvement and enhance their services.
- This project uses correlation analysis, association analysis, and k-means clustering to analyse airline reviews.
- The goal of the project is to provide insights that can help airlines improve their services and customer satisfaction, thereby boosting revenue.

Data Set

- The Airline Reviews Dataset is sourced from data.world, collected from Skytrax (https://www.airlinequality.com/).
- The dataset includes over 23,000 reviews gathered through web scraping, parsing HTML content, and organizing data.
- It contains around 19 attributes, including airline names, ratings, and detailed customer reviews.
- I choose this dataset because it has large text data and is ideal for machine learning and text analysis, offering meaningful insights into customer experiences.
- Original reviews from thousands of customers provide a comprehensive view of different airline service providers, useful for improving customer satisfaction.

Variable Name	Role	Type	Description
column_a	Id	Categorical	A unique ID for the review
airline_name	Feature	Categorical	Name of the airline
overall_rating	Feature	Integer	Overall rating
review_title	Feature	Categorical	Title of the review
review_date	Feature	Date	Date of the review posted
verified	Feature	Categorical	Verification status of the review
review	Feature	Categorical	Review
aircraft	Feature	Categorical	Type of the aircraft
type_of_traveller	Feature	Categorical	Type of the traveller
seat_type	Feature	Categorical	Type of the seat
route	Feature	Categorical	Route details
date_flown	Feature	Date	Date the passenger travelled
seat_comfort	Feature	Integer	Seat comfort rating
cabin_staff_service	Feature	Integer	Cabin staff rating
food_beverages	Feature	Integer	Food and beverages rating
ground_service	Feature	Integer	Ground services rating
inflight_entertainment	Feature	Integer	Inflight entertainment rating
wifi_connectivity	Feature	Integer	Wifi connectivity rating
value_for_money	Feature	Integer	Value for money Rating
recommended	Feature	Categorical	If flight can be recommended to other travellers

Data Cleaning

- All attributes were converted to their correct data types (text, integer, date) to ensure consistency and accuracy.
- Reviews with no airline name are were removed using Excel tools to maintain data integrity.
- Duplicate entries removed using the "Remove Duplicates" operator in RapidMiner.
- The "Replace Missing Values" operator from the RapidMiner is used to fill missing numerical values with the average value of their respective columns such as overall_rating, seat_comfort cabin_staff_service, food_beverages, ground_service, inflight_entertainment, wifi_connectivity and value_for_money, ensuring no gaps in critical data points.
- The "Filter Attribute" operator was used to exclude columns with missing values, focusing on complete data.
- The "Select Attribute" operator from the RapidMiner is used in selecting only the necessary or revelant attributes for the analysis.

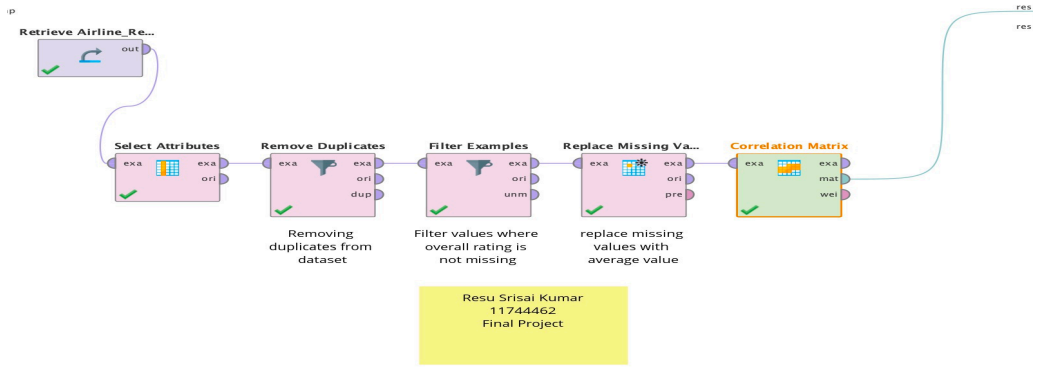
Goals

- **Relationship Analysis**: Investigate the relationships between various review attributes, such as:
  - Overall rating and seat comfort
  - Overall rating and inflight entertainment
- **Pattern and Association Discovery**: Identify patterns and associations within the reviews to gain insights into customer preferences and experiences.
- **Customer Segmentation**:
  - Group customers based on their reviews.
  - Help airline service providers identify different customer categories.
  - Tailor services to meet specific customer needs and preferences.
- This analysis aims to enhance the understanding of customer feedback, enabling airlines to improve their services and customer satisfaction.

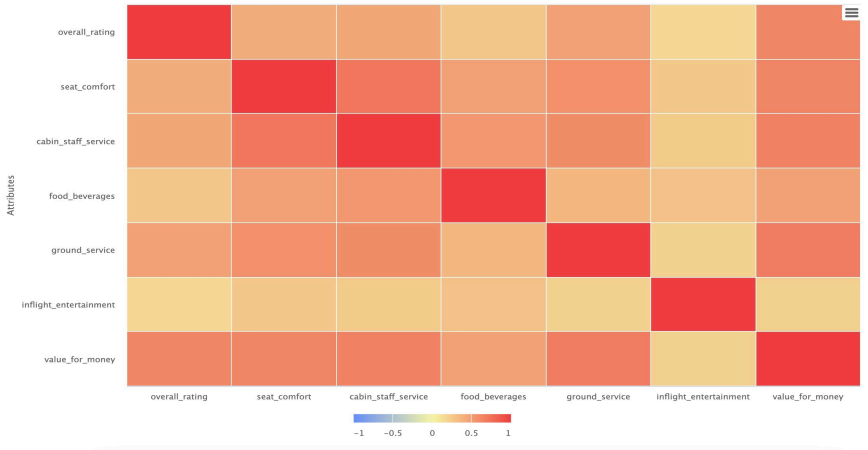
Data Analysis

Correlation Analysis

- Correlation between the numerical attributes are found from correlation analysis using rapid miner.
- Several components were used for data cleaning and data mining using RapidMiner
  - Select Attributes operator is used to select required numerical values.
  - Remove Duplicates operator is used to remove duplicate values.
  - Filter Examples operator is used to select rows which has review in it.
  - Replace Missing Values operator is used to replace missing numerical values with average value of the column.



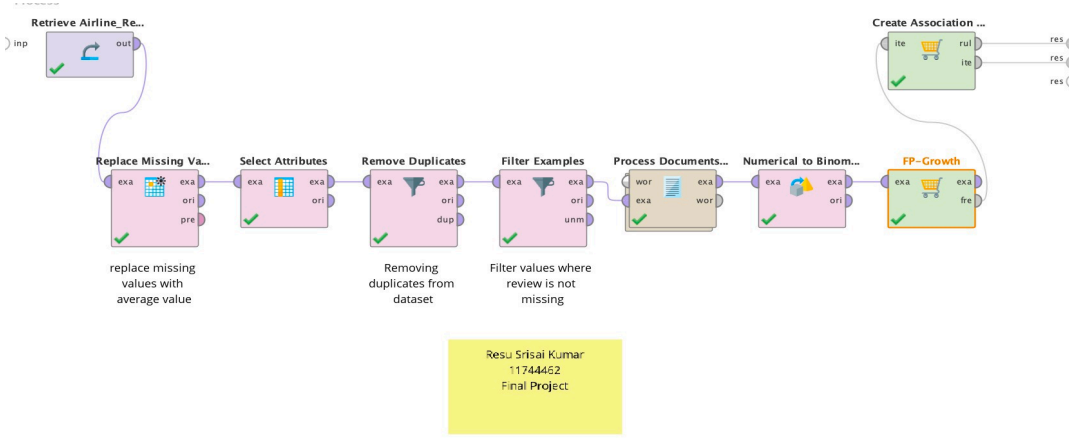
Attributes	overall_rating	seat_comfort	cabin_staff_service	food_beverages	ground_service	inflight_entertainment	value_for_money
overall_rating	1	0.390	0.420	0.251	0.461	0.159	0.602
seat_comfort	0.390	1	0.683	0.462	0.544	0.248	0.616
cabin_staff_service	0.420	0.683	1	0.508	0.561	0.237	0.622
food_beverages	0.251	0.462	0.508	1	0.333	0.283	0.456
ground_service	0.461	0.544	0.561	0.333	1	0.191	0.655
inflight_entertainment	0.159	0.248	0.237	0.283	0.191	1	0.214
value_for_money	0.602	0.616	0.622	0.456	0.655	0.214	1

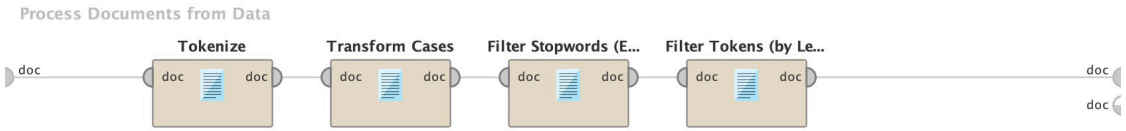


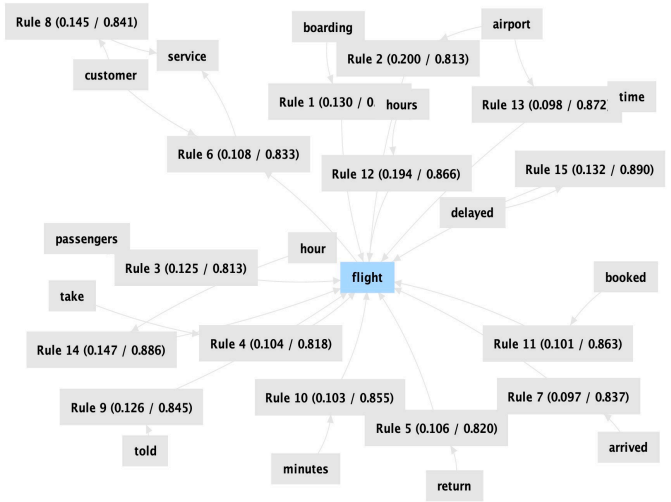
Data Analysis

Association Analysis

- Association analysis is used for developing the association or any similar patterns between the reviews.
- Similar components were used for data cleaning as in correlation analysis additionally
  - Tokenize operator is used to divide given text into words
  - Transform cases operator is used to convert words to lower letters
  - Filter stop words operator is used to remove stop words from the analysis
  - Filter tokens by length is used to include the words with specific length
- A numerical to binomial operator used to create association rules.



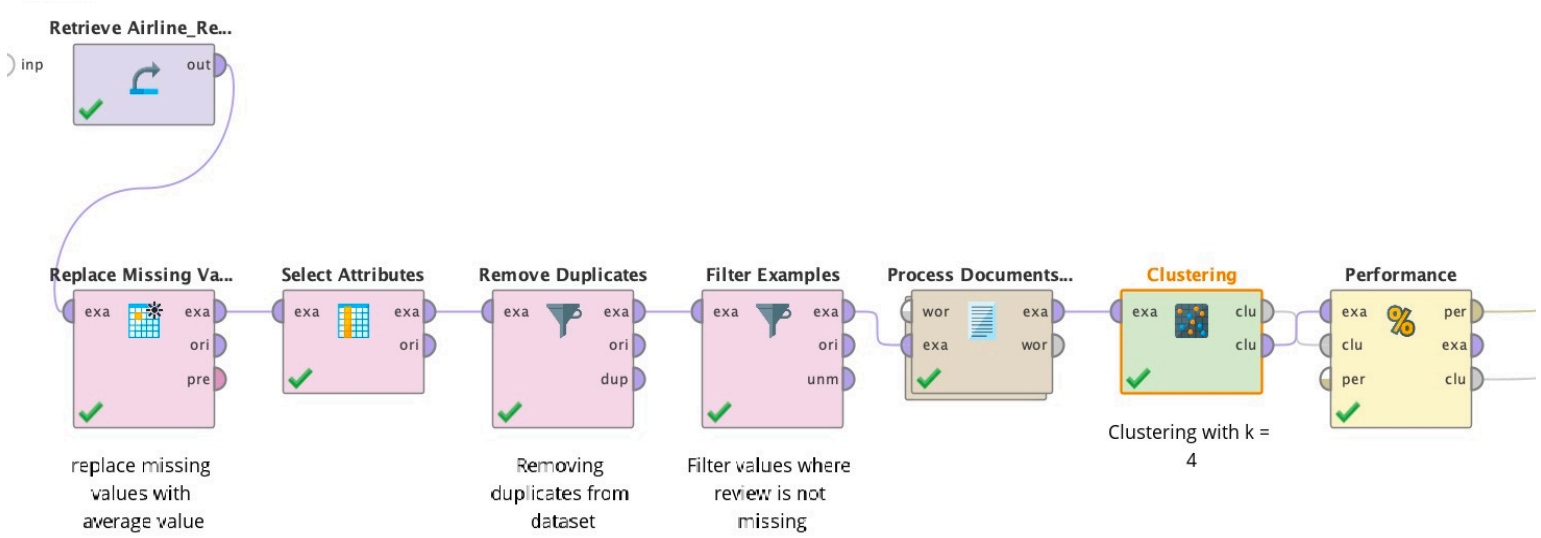


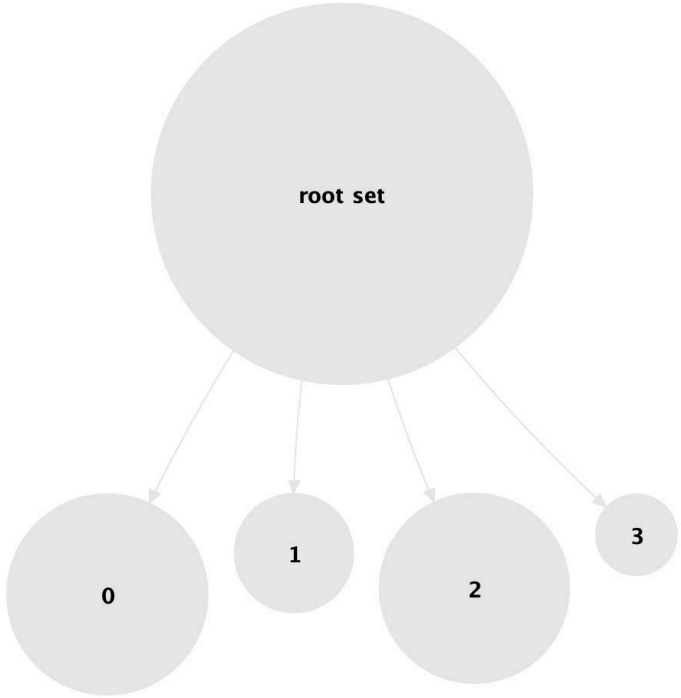


Data Analysis

K-Means Clustering Analysis

- K-means clustering is used to group similar reviews together based on their attributes which helps in discovering various hidden relations.
- K-Means Clustering analysis is performed using k = 4.





Resu Srisai Kumar
11744462
Final Project

Conclusion and Experience

This flight review analysis has revealed key insights:

- The analysis of airline reviews provided significant insights into customer experiences and preferences.
- Major issues such as flight delays and ticket cancellations were highlighted.
- Positive feedback included good inflight services and friendly staff.
- Strong correlations between service quality attributes and overall ratings were found, indicating key areas for improvement.
- This project gives valuable information from customer reviews which helps airline service providers better address customer needs and preferences.
- These findings can guide airlines in enhancing service quality and customer satisfaction.

Working on this project was a valuable learning experience. It enhanced my skills in data cleaning, correlation, clustering, and association analysis using RapidMiner which requires no code at all. Analyzing customer reviews provided meaningful insights that can help improve airline services.

Thank You