Titanic Association Analysis Project

Author: Anand Kumar Bhagat
Lamar University

Objective

The goal of this project is to analyze the Titanic dataset using **SAP Analytics Cloud (SAC)** and **Association Rule Mining** to understand passenger survivability based on various characteristics.

Dataset Overview

The dataset used for this analysis consists of Titanic passenger details from **Titanic.csv**, provided by the instructor. The analysis aims to identify association rules that highlight correlations between passenger class, gender, age group, and survival rates.

Software & Prerequisites

- **Software Used:** SAP Analytics Cloud (SAC)
- **Dataset Required:** Titanic.csv

Key Concepts in Association Rule Mining

- **Support:** Measures how frequently a rule or itemset appears in all transactions.

- **Confidence:** Indicates the frequency with which the consequent (RHS) occurs when the antecedent (LHS) is present.
- **Lift:** Measures how accurately a rule represents affinity or association compared to random co-occurrence of objects.

Methodology

1. **Importing and Preparing Data in SAC**

- Upload **Titanic.csv** into SAC.
- Ensure the **first row is set as column headers**.
- Modify **Passenger** aggregation to **Count**.
- Create a duplicate column of **Survived**, renaming it **RHS**.
- Concatenate **Class, Sex, and Age** to create a new column **LHS**.
- Further concatenate **LHS** and **RHS** to generate **Rules**.

---

2. Identifying the Rules

- **Most Frequent Rule:** "Crew, Male, Adult, No" appeared most frequently.
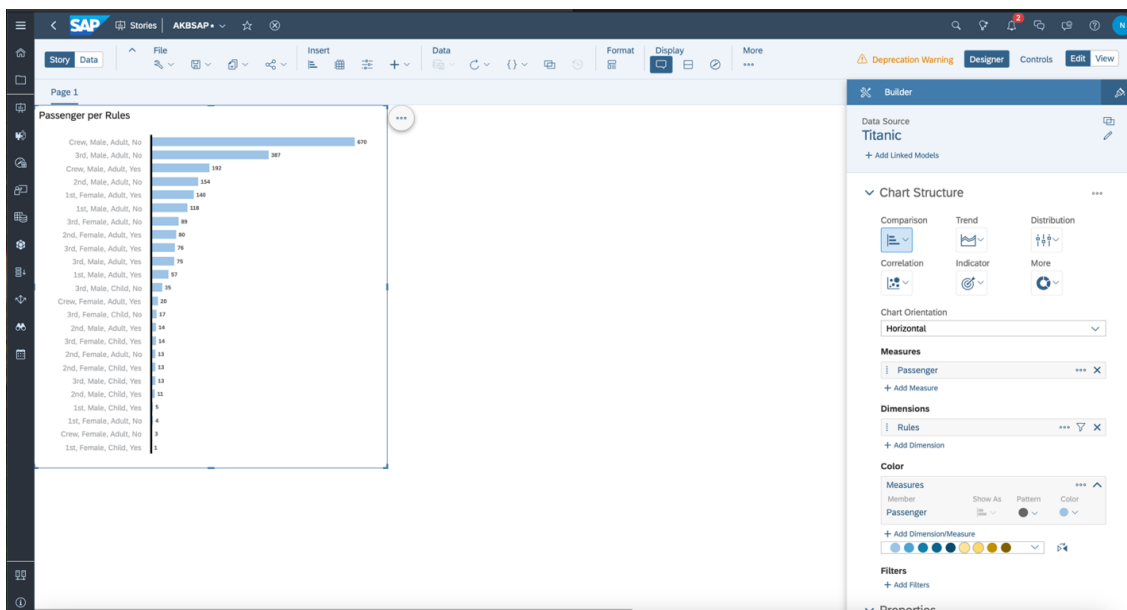
The Bar Chart below represents the most frequently occurring rule in the data set. The rule "Crew, Male, Adult, No" appears the most times, indicating that many crew members who were adult males did not survive.

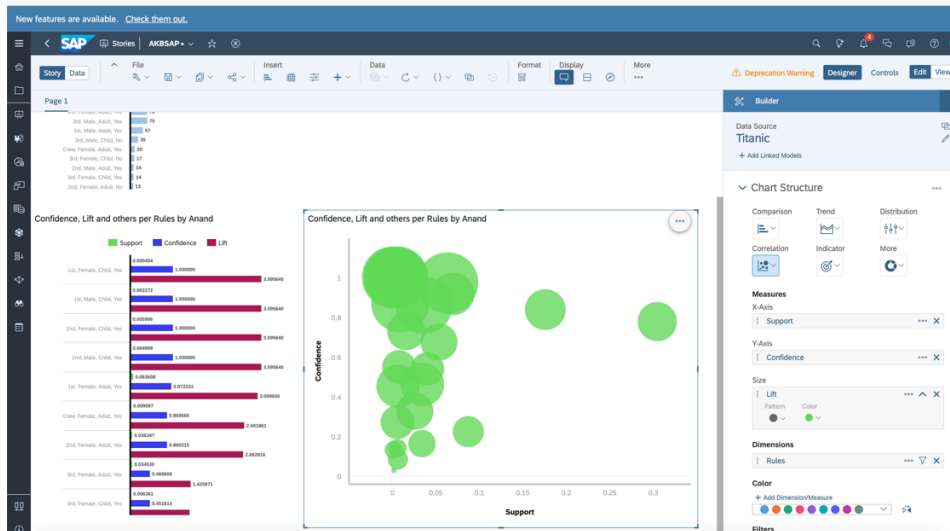- **Strongest Rule:** "2nd Male Child, Yes" had the highest **Lift**.

Strongest Rule:

☐ **Explanation: The strongest rule, "2nd Male Child, Yes", has the highest Lift value, indicating that second-class male children had the highest probability of survival compared to other groups.**
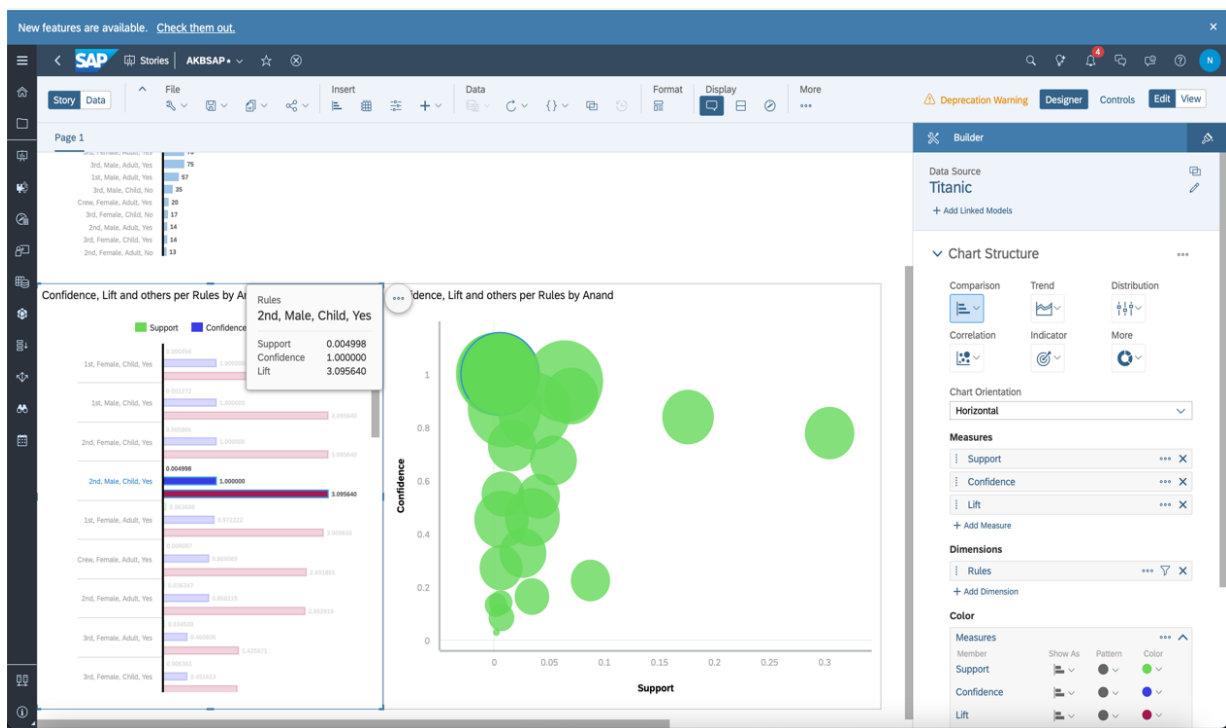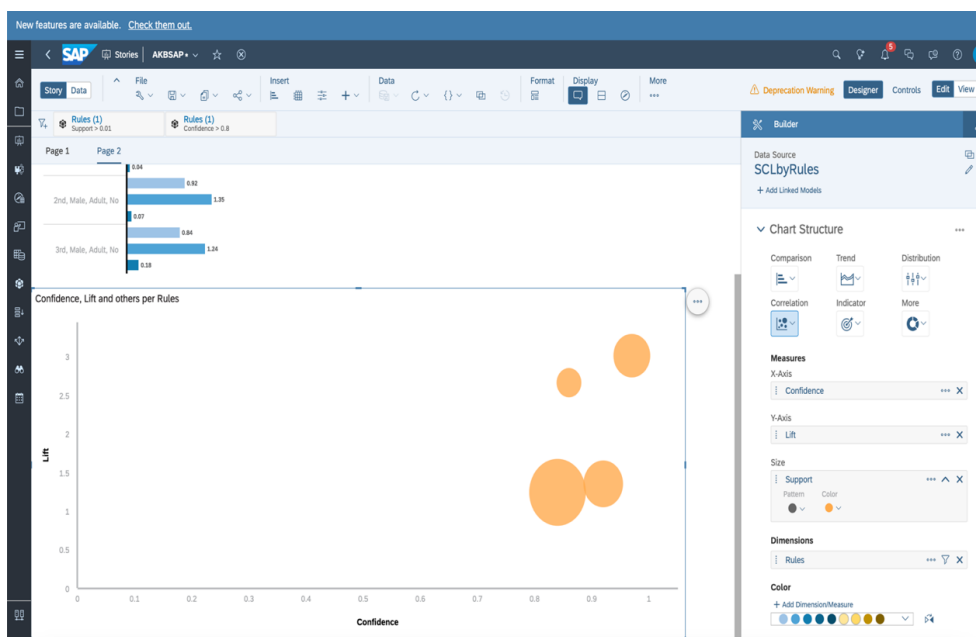
Creating Visualizations:

Bar Chart: Displays the frequency of different rules.

Bubble Chart: Represents Support, Confidence, and Lift / Filtered Chart: Helps refine the most important rules using Confidence & Support.
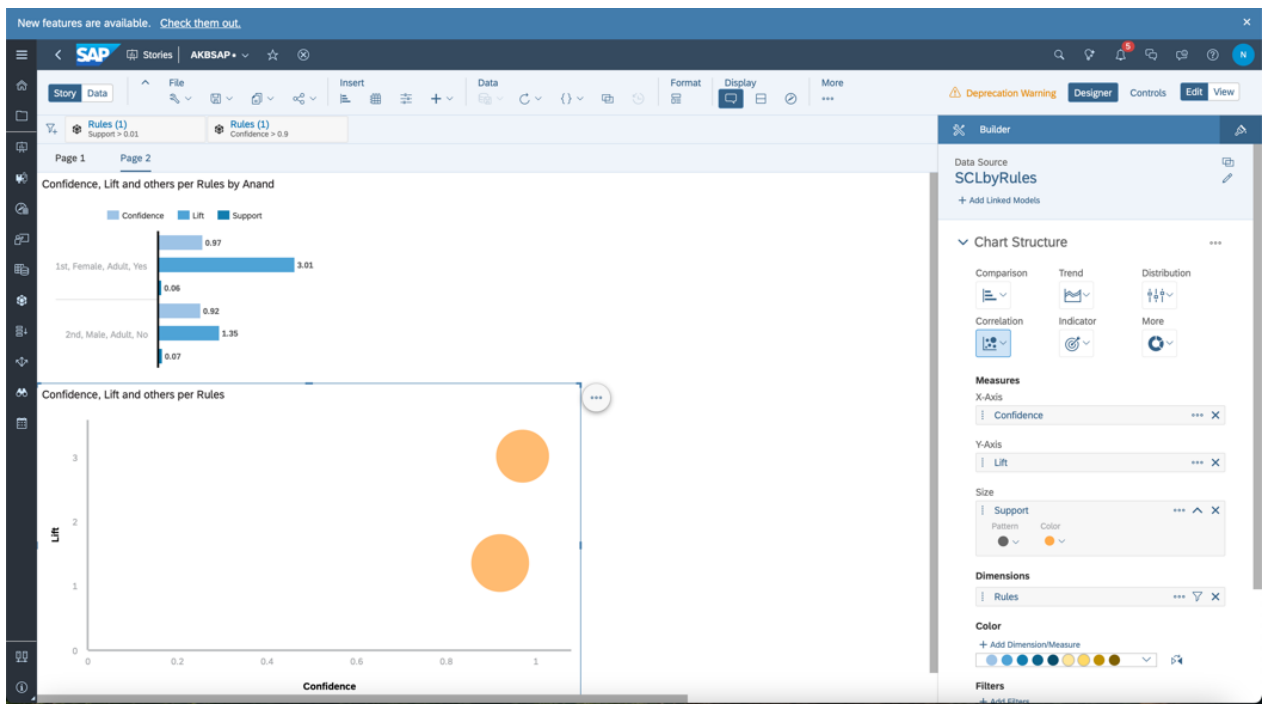
Effect of Increasing Confidence and Support on Association Analysis.

Explanation: Increasing both Confidence and Support values in the Association Analysis filters will reduce the number of rules generated. This happens because the filters set stricter thresholds, allowing only the most frequent and reliable associations to be included. The result is a more focused analysis with fewer, stronger relationships that are easier to interpret.

Steps:

- Navigate to your Association Analysis chart or table.

- Set the **Support filter** to a higher value (e.g., above **0.01**).

- Set the **Confidence filter** to a higher value (e.g., above **0.80**).

- Apply the updated filters to your analysis.

- Notice that fewer rules will be displayed because only the stronger, more significant associations remain.

- **Support:** Greater than 0.01
- **Confidence:** Greater than 0.80

Conclusions/Results:

- By increasing both Confidence and Support, the number of rules decreases, leaving only the most meaningful and strongest associations.
- The analysis becomes clearer and more focused, as weak or irrelevant rules are filtered out.
- The visualization will be less crowded, and the remaining rules are more significant, providing actionable insight.

Effect of Decreasing Confidence and Support.

 Explanation:

When **both Confidence and Support values are decreased**, more association rules appear in the analysis. However, many of these rules may not be meaningful, leading to an **overcrowded visualization** with weaker relationships.

Steps to Perform in SAP Analytics Cloud (SAC):

1. **Lower the Support value** (e.g., set it below 0.01).
2. **Lower the Confidence value** (e.g., set it below 0.50).
3. **Apply the filters** and observe how many additional rules are displayed
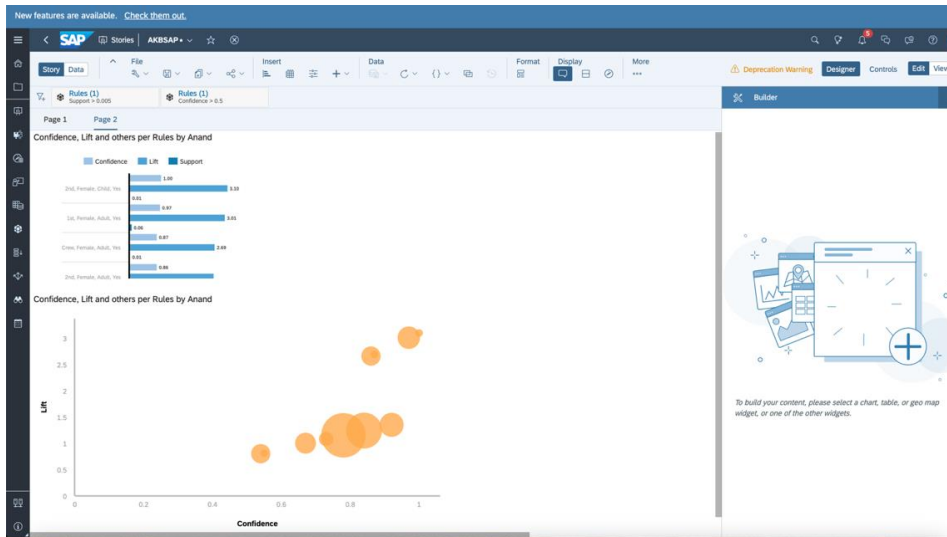
□ **New Filter Values Used:**

- **Support:** Less than 0.01
- **Confidence:** Less than 0.50

Conclusion:
Lowering these values increases the number of generated rules, but it also makes it harder to identify strong and reliable associations. Many of the rules may be weak, leading to an excessive amount of information that is difficult to interpret.

Visual Representations

- **Bar Chart:** Displays the most frequently occurring rules.
- **Bubble Chart:** Highlights associations with varying degrees of support, confidence, and lift.
- **Filtered Bubble Chart:** Shows only the strongest rules after applying filters.

(Screenshots from the analysis should be placed in these sections.)

Conclusion

This association analysis provided valuable insights into survival patterns on the Titanic. Key findings include:

- **Crew members, particularly adult males, had a lower survival rate**.
- **Young children in second-class had a high likelihood of survival**.
- **Filtering by Support and Confidence helps identify the most relevant rules**.

Next Steps

- Publish this project on LinkedIn and Medium to showcase data analytics expertise.
- Explore other machine learning techniques on the Titanic dataset, such as classification models.
- Apply similar association analysis techniques to different datasets for deeper insights.

Author: Anand Kumar Bhagat
Portfolio: https://anandlamar.github.io/