**ALTERNATIVE ASSESSMENT 1**

——Case Study: E-Commerce Customer Behaviour Analysis

1. **Dataset Overview: E-commerce Customer Behavior**

* Description

This dataset offers a holistic perspective on customer behavior within an e-commerce platform. Each entry corresponds to a unique customer, providing an intricate breakdown of interactions and transactions. The dataset is meticulously designed to facilitate nuanced analysis, offering insights into customer preferences, engagement patterns, and satisfaction levels. This comprehensive view empowers businesses to make informed, data-driven decisions aimed at enhancing the overall customer experience.

* Columns
* Customer ID

Type: Numeric

Description: A unique identifier assigned to each customer for dataset distinctiveness.

* Gender

Type: Categorical (Male, Female)

Description: Specifies the gender of the customer, enabling gender-based analytics.

* Age

Type: Numeric

Description: Represents the age of the customer, facilitating age-group-specific insights.

* City

Type: Categorical (City names)

Description: Indicates the city of residence for each customer, providing geographic insights.

* Membership Type

Type: Categorical (Gold, Silver, Bronze)

Description: Identifies the type of membership held by the customer, influencing perks and benefits.

* Total Spend

Type: Numeric

Description: Records the total monetary expenditure by the customer on the e-commerce platform.

* Items Purchased

Type: Numeric

Description: Quantifies the total number of items purchased by the customer.

* Average Rating

Type: Numeric (0 to 5, with decimals)

Description: Represents the average rating given by the customer for purchased items, gauging satisfaction.

* Discount Applied

Type: Boolean (True, False)

Description: Indicates whether a discount was applied to the customer's purchase, influencing buying behavior.

* Days Since Last Purchase

Type: Numeric

Description: Reflects the number of days elapsed since the customer's most recent purchase, aiding in retention analysis.

* Satisfaction Level

Type: Categorical (Satisfied, Neutral, Unsatisfied)

Description: Captures the overall satisfaction level of the customer, providing a subjective measure of their experience.

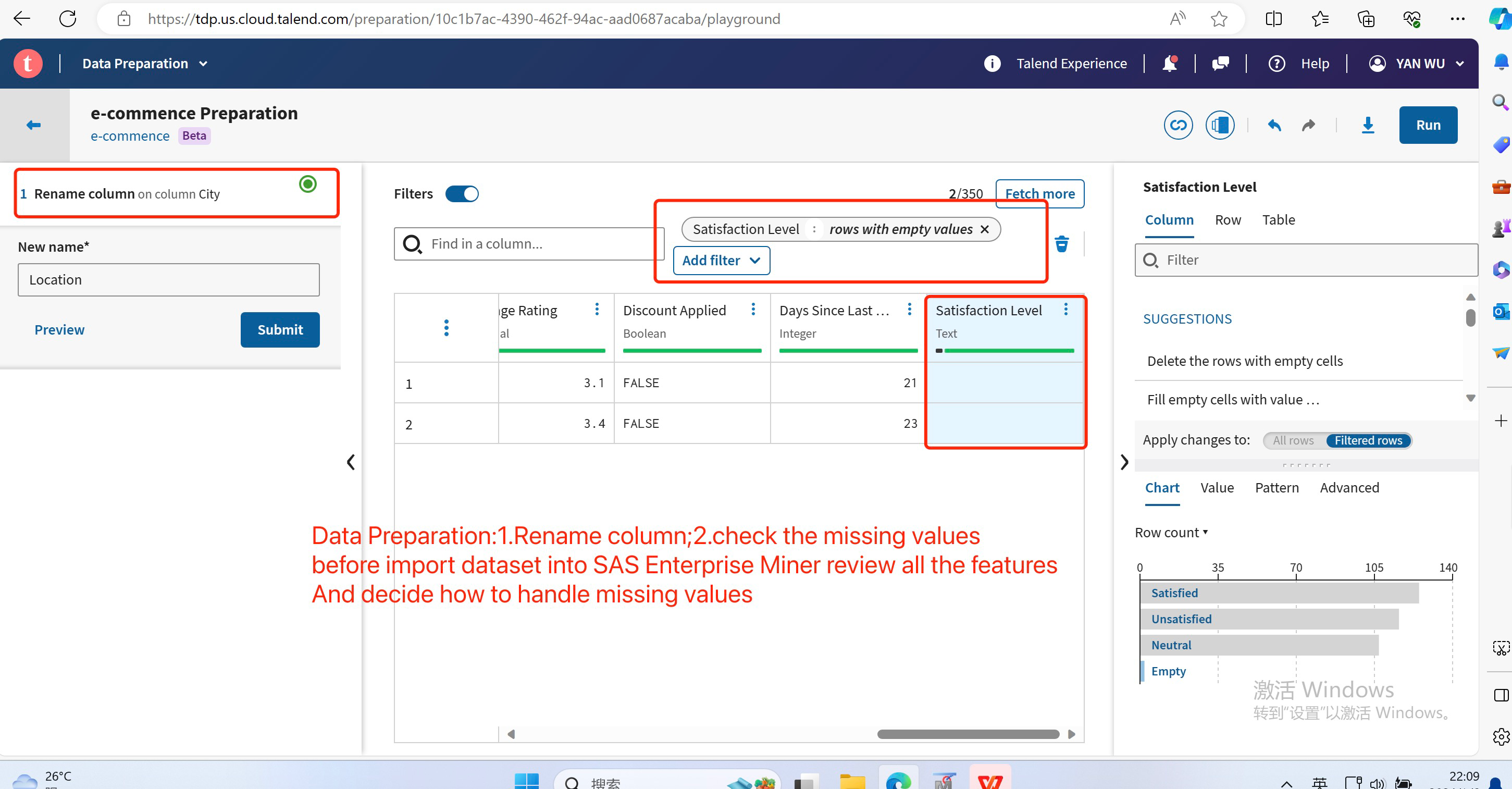
This dataset, with its rich attributes, lays the foundation for a profound exploration of e-commerce customer behavior. Analyzing these dimensions promises valuable insights for strategic decision-making and customer-centric improvements.

1. **Utilizing Talend for initial data handling**

In this phase, I performed several key actions using Talend to align the dataset with the requirements of AA1 and cater to the specifics of our case study, "E-Commerce Customer Behaviour Analysis." Notably, **I opted to rename the feature "City" to "Location" for closer alignment with AA1 and enhanced relevance to our analysis.** This modification ensures that the dataset is more attuned to the nuances of our specific analytical goals.

**Additionally, I refrained from transforming the "Days Since Last Purchase" feature into "LastPurchaseDate: The date of the last purchase."** The decision was rooted in the lack of explicit information about the transaction period within the dataset. Converting this feature might introduce uncertainties into our analysis rather than aiding it.

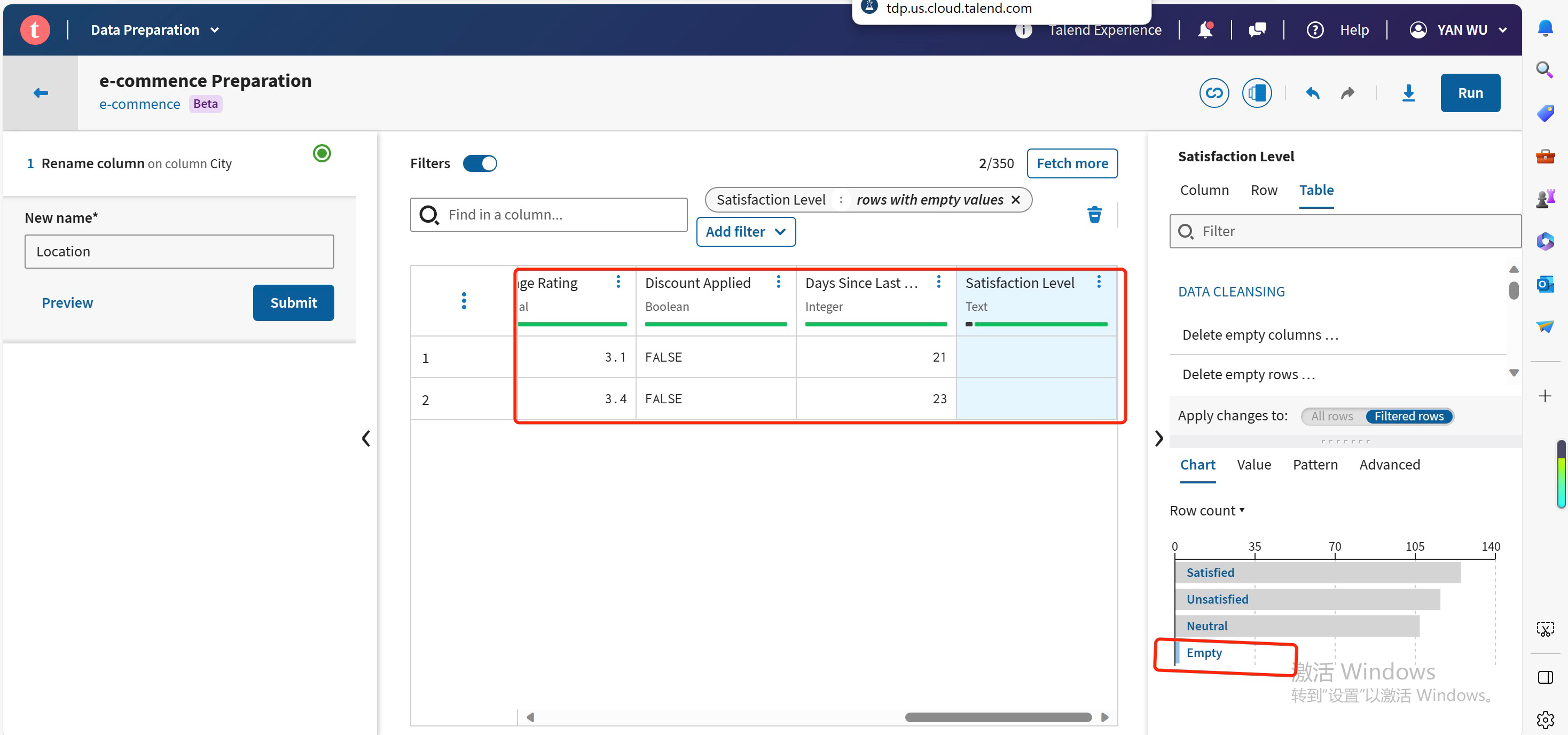
Furthermore, I meticulously examined each feature in the dataset using Talend Data Preparation to assess data quality. Identifying the precise locations of missing values enabled a targeted strategy for handling these gaps. Leveraging Talend Data Preparation, I analyzed and selected appropriate strategies for addressing missing values, laying a robust foundation for SAS Enterprise Miner to conduct thorough and efficient data analysis. This meticulous approach ensures that the subsequent stages of analysis are built upon a clean, well-structured dataset.

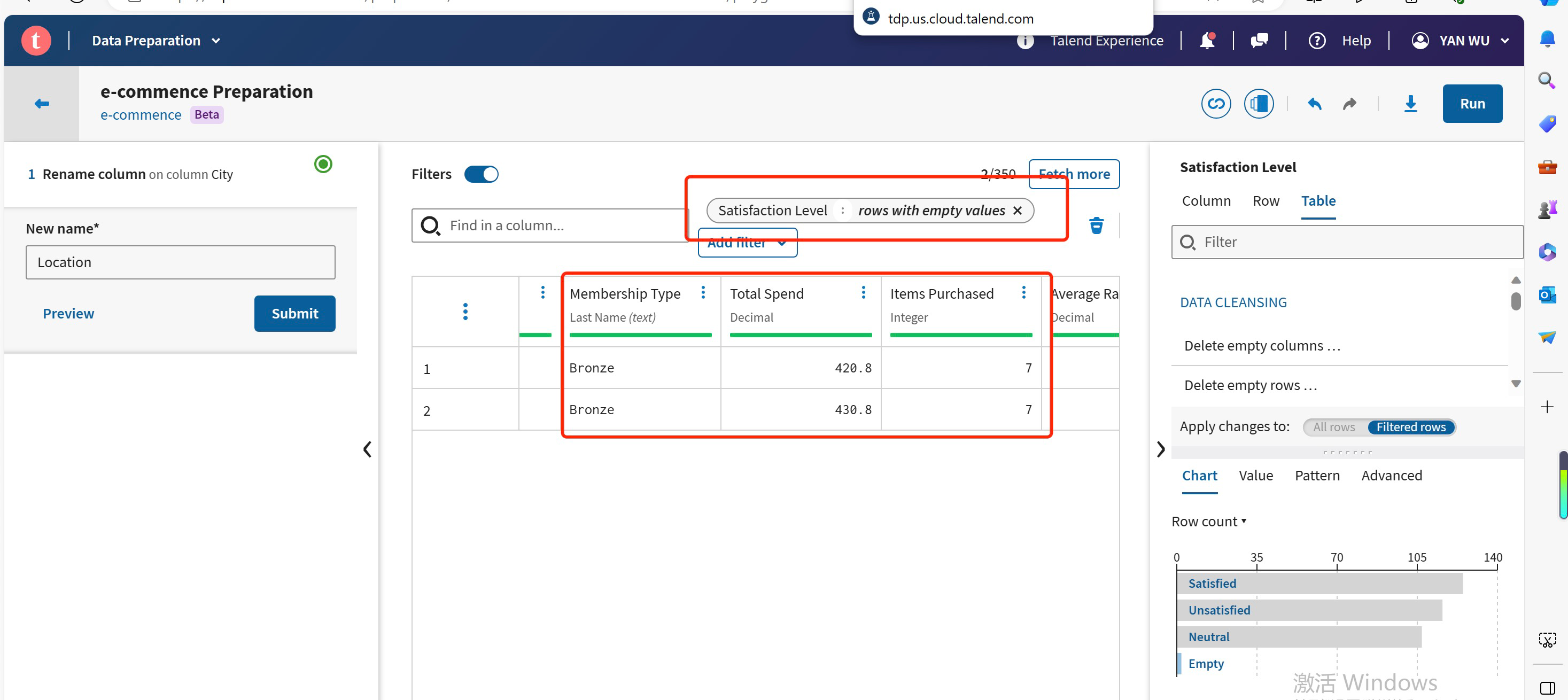


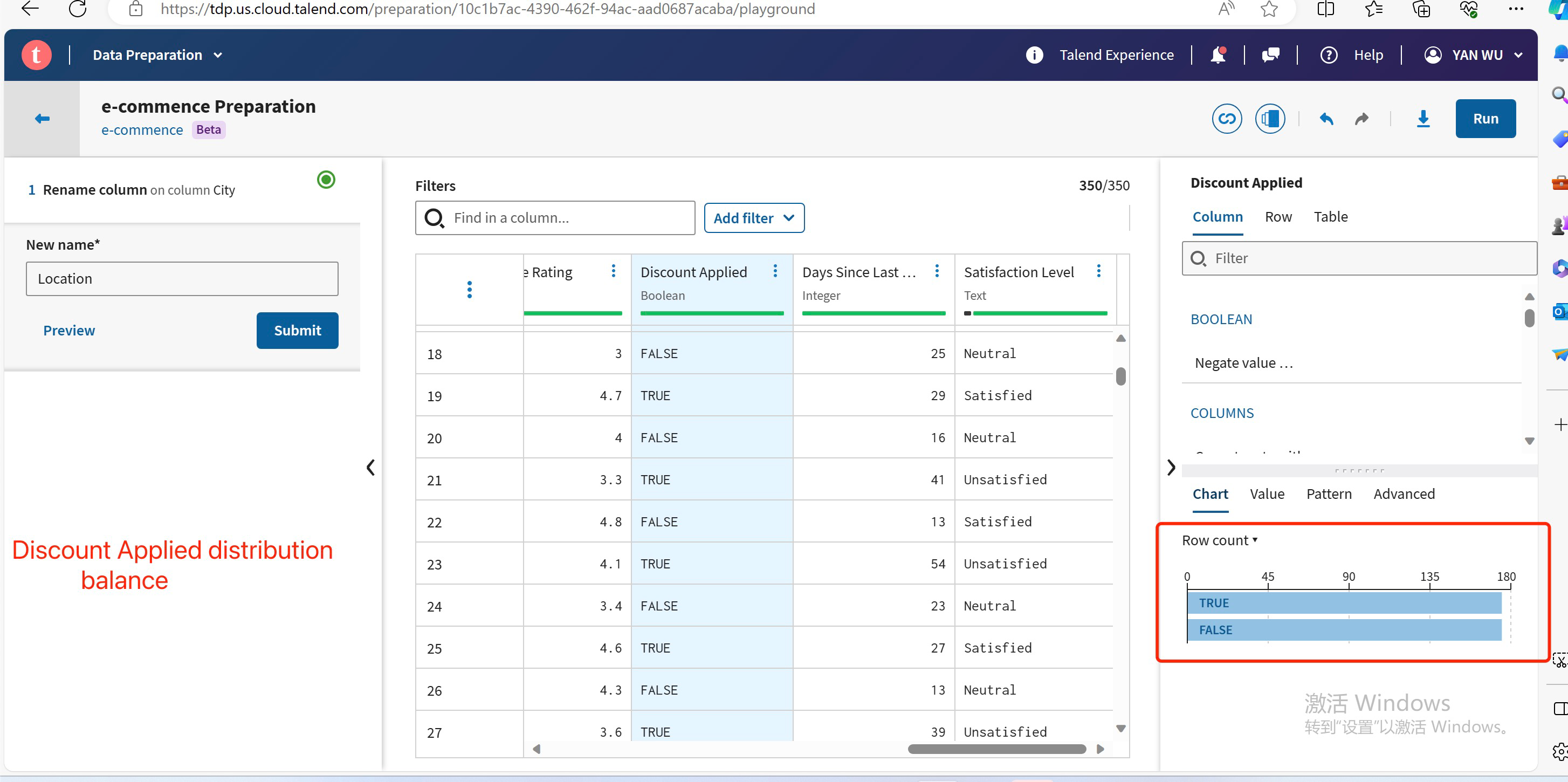
1. **Select Stratagy to handle Missing Values in the "Satisfaction Level" Column**

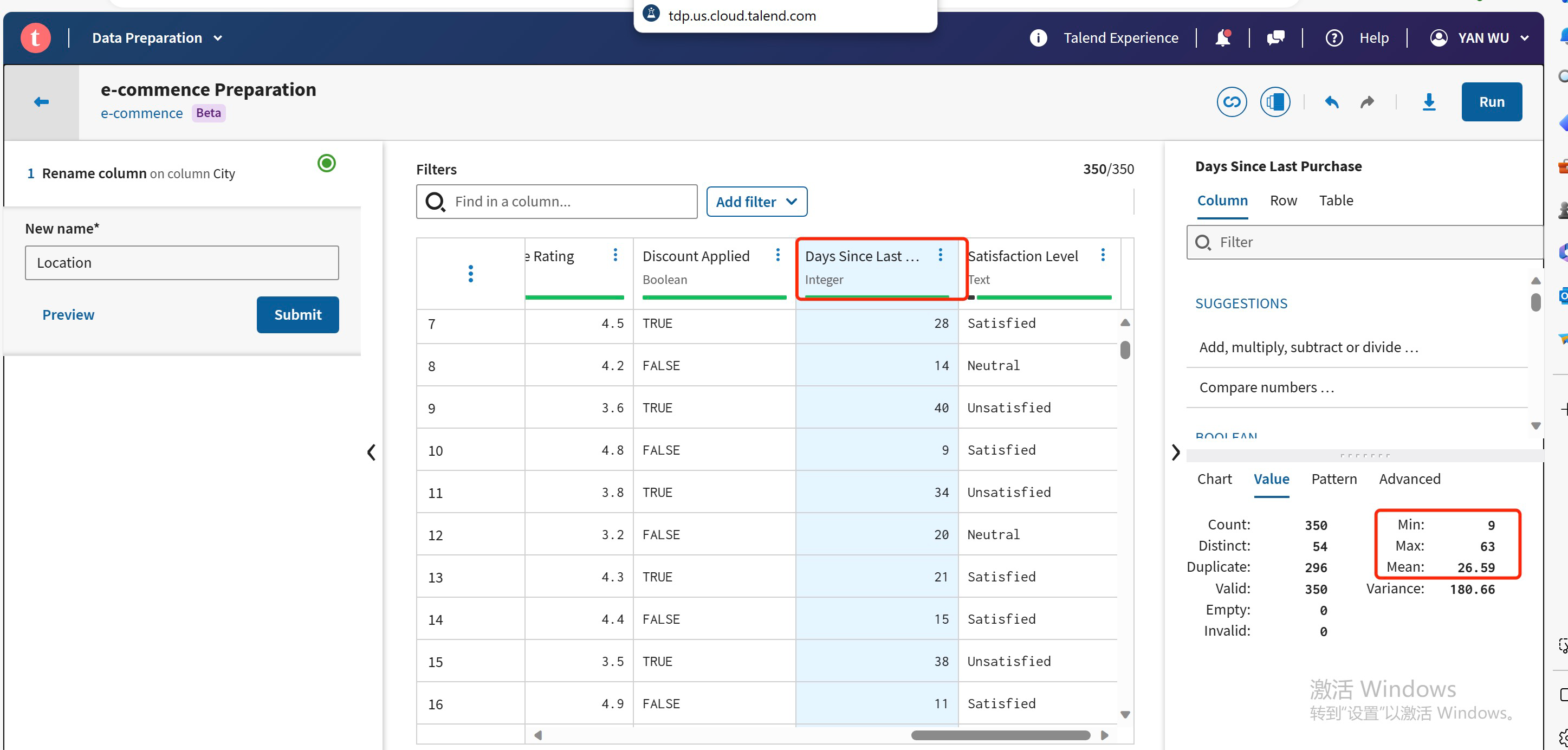
There are **two** missing values in the "Satisfaction Level" column, denoted by empty cells. Given the small number of missing values and their occurrence in the "Satisfaction Level" column, we need to devise a strategy for imputing these values based on business logic.

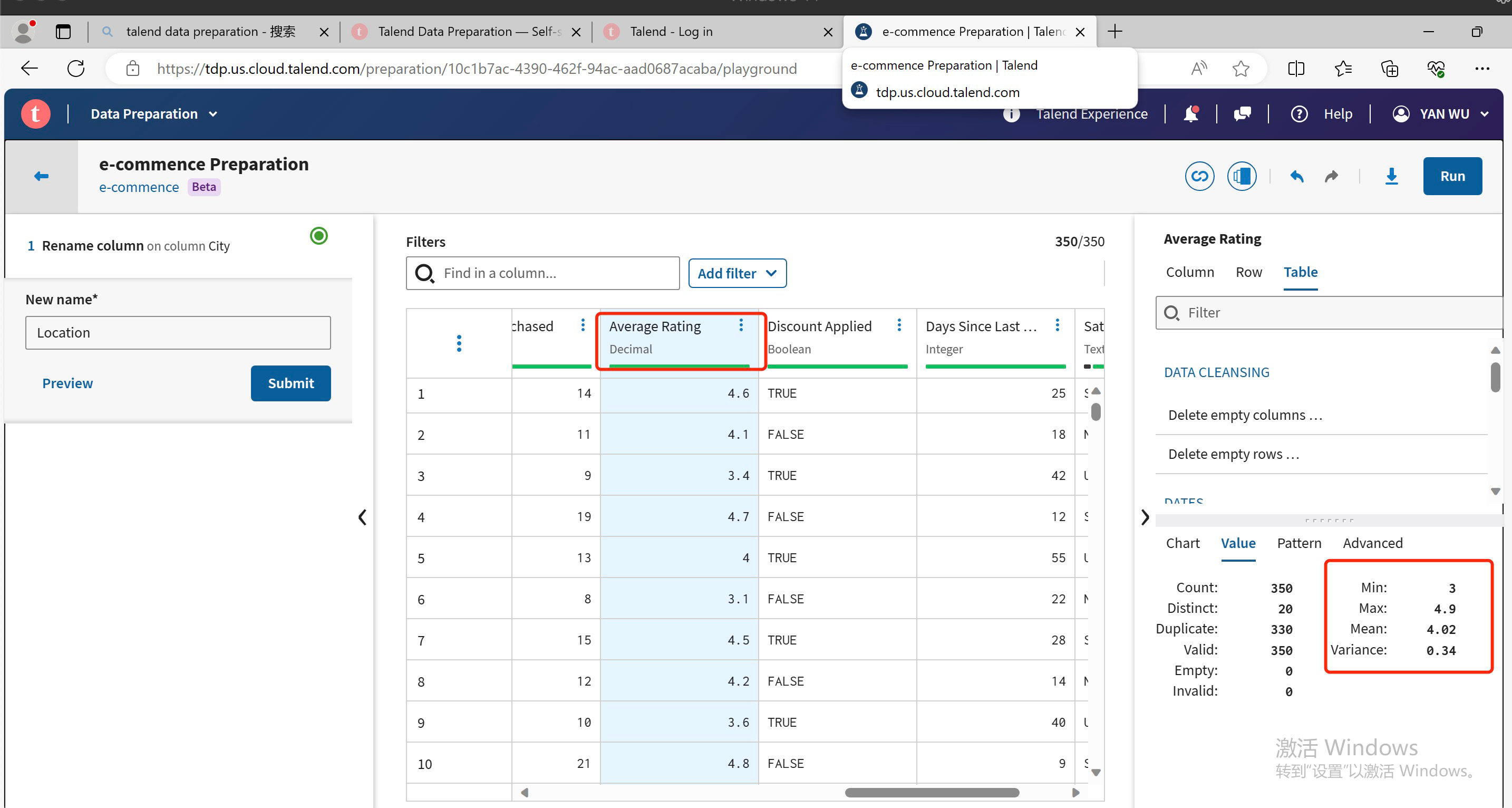
Upon reviewing related features, the "Days Since Last Purchase" column stands out as potentially influential. Analyzing its statistical information reveals that the maximum, minimum, and average values are 63 days, 9 days, and 26.59 days, respectively. Interestingly, the missing values in the "Satisfaction Level" column correspond to "Days Since Last Purchase" values of 21 days and 23 days, both below the average. Additionally, the distribution of the "Discount Applied" feature is evenly split between TRUE and FALSE.



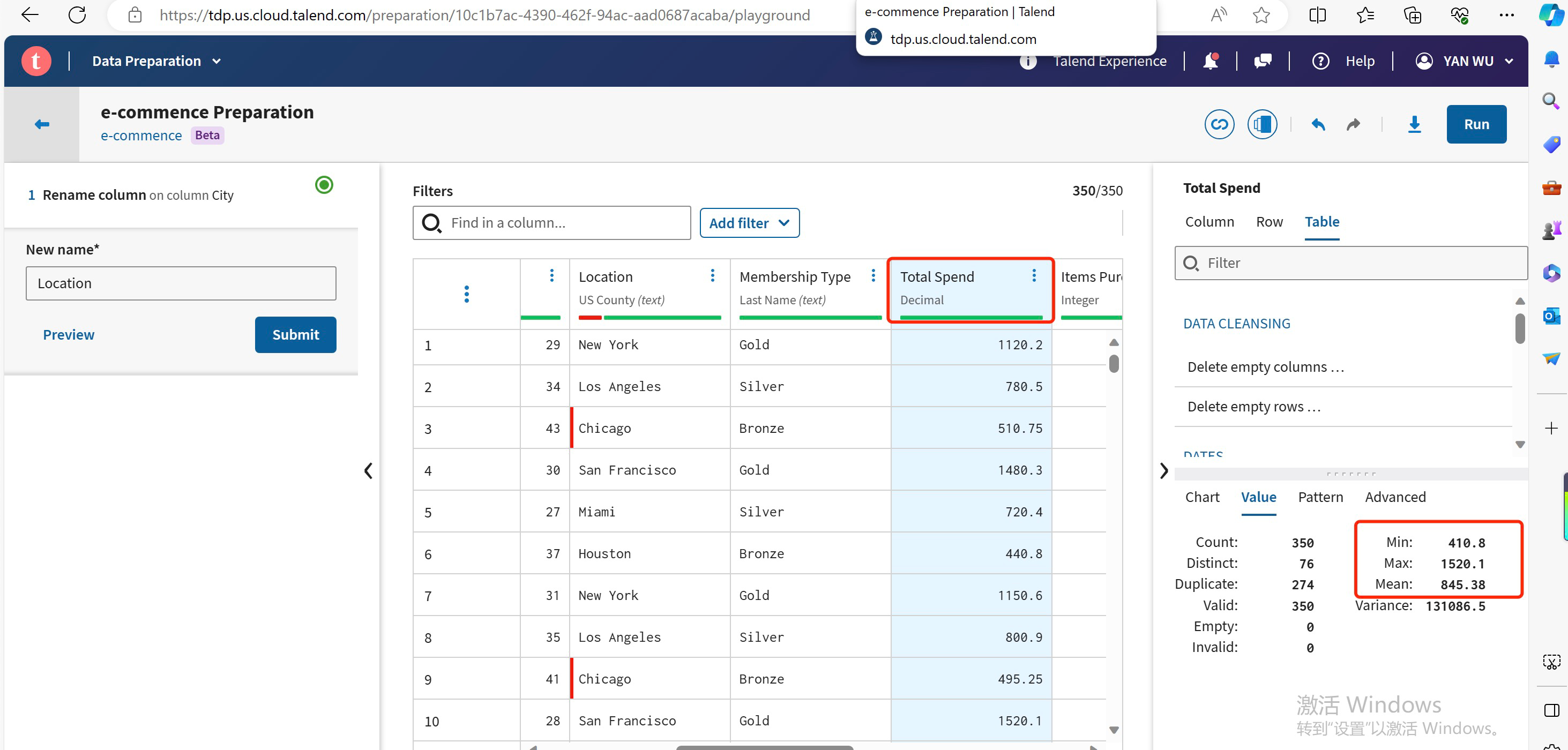












Considering the interplay of these features, it is plausible that the instances with missing values **represent new customers on the e-commerce platform.** Their satisfaction levels are crucial for our customer behavior analysis. **Eliminating these instances (Remove Rows with Missing Values) is not desirable** due to the potential loss of valuable information. Instead, I opt for an approach grounded in business logic.

* Imputation Strategy

I decide to **impute the missing values in the "Satisfaction Level" column with a neutral value**, reflecting a neutral satisfaction level. This decision is substantiated by the characteristics of the instances with missing values. The associated "Days Since Last Purchase" values suggest relatively recent engagements, while other features indicate a Bronze membership, lower total spend, and fewer items purchased—attributes often indicative of new customers.

By imputing with a neutral value,I acknowledge the absence of explicit satisfaction feedback from these new customers while introducing a realistic and neutral assumption.

1. **Data Import and Preprocessing by SAS Enterprise Miner**

In this step, the focus shifts to SAS Enterprise Miner for the importation and preprocessing of our dataset. The following actions were undertaken:

* Dataset Importation

The dataset, refined through initial handling in Talend, was seamlessly imported into SAS Enterprise Miner.

This step ensures a smooth transition of the dataset for further in-depth analysis and modeling within the SAS environment.

* Handling Missing Values

Leveraging the strategies defined during the Talend phase, missing values were addressed systematically within SAS Enterprise Miner.

Imputation based on business logic was applied, especially in critical features like "Satisfaction Level," where missing values were replaced with the category "Neutral."

* Variable Role Specification

Variable roles were meticulously specified to align with the analytical goals of the case study.

* Customer ID

Role Assigned: Input

Rationale: As a unique identifier, "Customer ID" is designated as an input variable to maintain dataset distinctiveness. This variable doesn't serve as a target or predictor but is essential for maintaining individuality.

* Gender

Role Assigned: Input

Rationale: "Gender" is assigned the role of an input variable, allowing for gender-based analytics. This information can provide valuable insights into spending patterns and preferences.

* Age

Role Assigned: Input

Rationale: Designated as an input variable, "Age" facilitates age-group-specific insights. Understanding the age distribution is crucial for tailoring marketing strategies.

* Location

Role Assigned: Input

Rationale: "Location" is assigned the role of an input variable to provide geographic insights. Regional variations in customer behavior can significantly impact marketing efforts.

* Membership Type

Role Assigned: Input

Rationale: Designated as an input variable, "Membership Type" influences perks and benefits. Analyzing this variable can help tailor marketing strategies based on membership levels.

* Total Spend

Role Assigned: **Target**

Rationale: Chosen as the target variable for Customer Segmentation, "Total Spend" aligns with the objective of analyzing and categorizing customers based on their spending habits. This variable is crucial for marketing and promotional campaigns.

* Items Purchased

Role Assigned: Input

Rationale: "Items Purchased" is designated as an input variable, providing insights into the quantity of items bought by customers. This information contributes to understanding shopping behaviors.

* Average Rating

Role Assigned: Input

Rationale: Assigned the role of an input variable, "Average Rating" represents customer satisfaction. It contributes to the overall understanding of customer sentiment.

* Discount Applied

Role Assigned: Input

Rationale: Designated as an input variable, "Discount Applied" indicates whether a discount influenced purchasing behavior. This variable is crucial for assessing the impact of discounts on customer spending.

* Days Since Last Purchase

Role Assigned: Input

Rationale: Assigned as an input variable, "Days Since Last Purchase" aids in retention analysis by reflecting the time gap since a customer's most recent purchase.

* Satisfaction Level

Role Assigned: Input

Rationale: Although representing overall satisfaction, "Satisfaction Level" is designated as an input variable. This decision is made to avoid multicollinearity issues, considering "Total Spend" as the target variable.

This meticulous assignment of variable roles ensures a comprehensive understanding of customer behavior, enabling effective segmentation for targeted marketing strategies.

1. **Decision Tree Analysis in SAS Enterprise Miner**

* Advanced Exploratory Data Analysis (EDA)
* Correlation Analysis

Right-click on the dataset, select "Advanced," and choose "Correlation Matrix."

Investigate correlations between variables, looking for strong positive/negative correlations.

* Univariate Analysis

Utilize the "Distribution Analysis" node for detailed univariate analysis.

Explore individual columns for outliers, patterns, or specific characteristics.

* Bivariate Analysis

Combine the "Distribution Analysis" node with the "Explore" node for bivariate analysis.

Explore relationships between pairs of variables to uncover dependencies.

* Multivariate Analysis

Integrate the "Principal Components" node for dimensionality reduction and multivariate analysis.

Identify complex dependencies and patterns among multiple variables.

Let's go step by step for the Advanced EDA in SAS Enterprise Miner:

* Customer ID - Unique Identifier and Checking Duplicates Values

Use the "Explore Data" node in SAS Enterprise Miner.

Drag the "Customer ID" variable to the "Unique Identifier" role.

Explore the data to check for any duplicate values.

Rationale:

Ensures each customer is uniquely identified.

Detects and handles any duplicate entries.

* Gender - Analyzing Distribution and Exploring Patterns

Use the "Explore Data" node.

Drag the "Gender" variable to the "Analysis" role.

Examine distribution charts and statistics.

Rationale:

Understand the gender distribution in the dataset.

Identify any patterns or trends related to gender.

* Age - Analyzing Distribution and Investigating Trends

Continue using the "Explore Data" node.

Drag the "Age" variable to the "Analysis" role.

Explore distribution and trends.

Using Recode Node:

Drag and drop a "Recode" node onto the diagram workspace.

Connect the input dataset to the "Recode" node.

Configure Recode Node:In the "Recode" node, select the variable (e.g., Age) that want to recode.

Define the recoding rules using the interface. Create bins and assign labels.

Execute the "Recode" node to apply the recoding rules and create a new variable for age groups.

Rationale:

Gain insights into the age distribution.

Investigate trends or patterns in customer ages.

* Location - Exploring Spending Patterns and Segmentation

Utilize the "Explore Data" node.

Drag the "Location" variable to the "Analysis" role.

Investigate spending patterns and potential segmentation based on location.

Rationale:

Identify regions with specific spending behaviors.

Consider location-based segmentation strategies.

* Membership Type - Analyzing Distribution and Correlations

Use the "Explore Data" node.

Drag the "Membership Type" variable to the "Analysis" role.

Explore distribution and check for correlations with other variables.

Rationale:

Understand the distribution of membership types.

Identify relationships between membership type and other variables.

* Visualization

Visualize key insights using SAS Enterprise Miner's visualization capabilities.

Create charts or graphs to represent patterns identified during EDA.