Talend Data Prep Steps

Beginning Exploration with Talend Data Prep.
For the beginning of the exploration, the original data was first imported into Talend Data Prep for a quick glance on the key attributes and to work on generating extra columns that is more suitable for the case study.

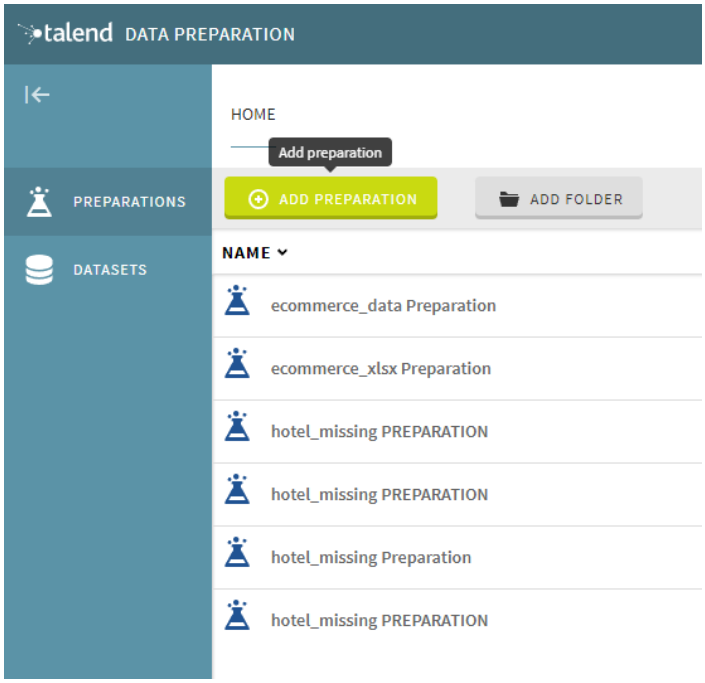


Figure 1: Click Add Preparation

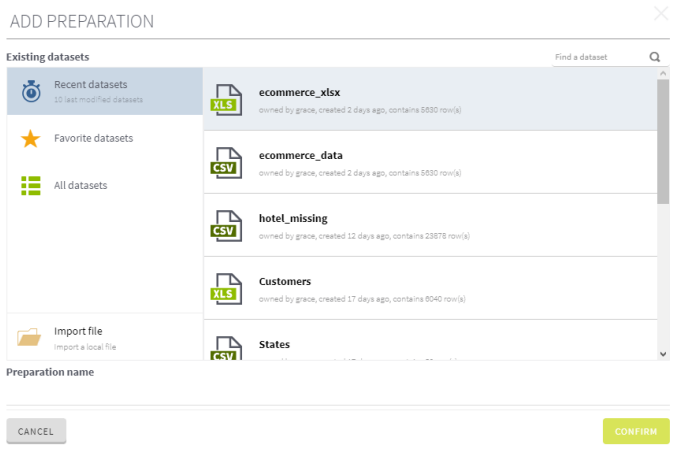


Figure 2: Select ecommerce_data

The original dataset was the ecommerce_xlsx. After checking the advanced charts for some of the key values, I was able to determine the conditions needed to generate the MembershipLevel

variable. Then, I used the conditions I identified using Talend Data Prep and generated the dataset using JupyterNotebook.

With the final dataset generated in Jupyternotebook, I then imported the official dataset into Talend Data Prep to do more exploration. Refer to the step in figure 1 and figure 2 to import dataset into Talend Data Prep.

Exploration of Dataset Using Quality Bars, Charts, Value, and Patterns

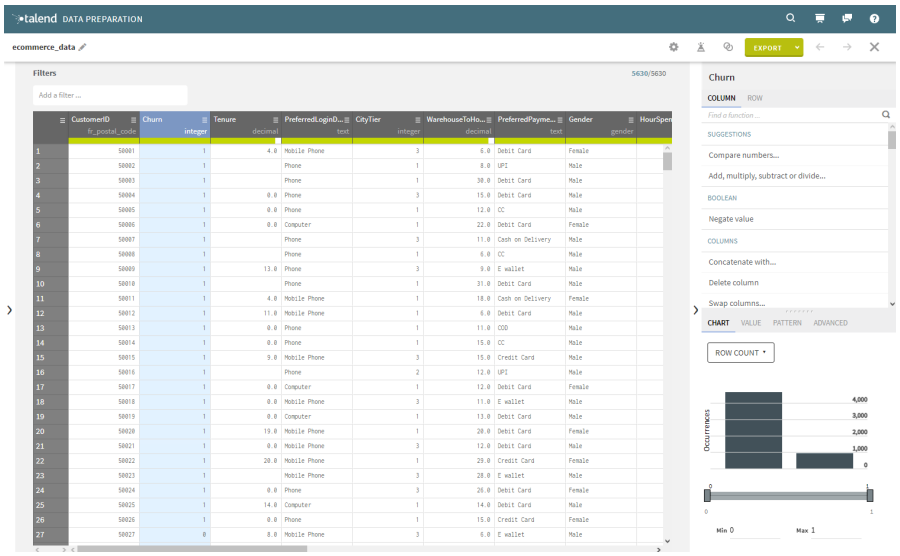


Figure 1: Data Profiling in Talend Data Prep (Part 1)

	HourSpendOnApp decimal	NumberOfDevic... integer	PreferedOrderCat text	SatisfactionScore integer	MaritalStatus text	NumberOfAddress integer	Complain integer	OrderAmountHi... decimal
6	3.0	5	Mobile Phone	5	Single	2	1	22.0
7	2.0	3	Laptop & Accessory	2	Divorced	4	0	14.0
8	3.0	3	Mobile	2	Divorced	3	1	16.0
9		4	Mobile	3	Divorced	2	1	14.0
10	2.0	5	Mobile	3	Single	2	0	12.0
11	2.0	3	Others	3	Divorced	2	0	
12	3.0	4	Fashion	3	Single	10	1	13.0
13	2.0	3	Mobile	3	Single	2	1	13.0
14	3.0	4	Mobile	3	Divorced	1	1	17.0
15	3.0	4	Fashion	2	Single	2	0	16.0
16	3.0	3	Mobile	5	Married	5	1	22.0

Figure 2: Data Profiling in Talend Data Prep (Part 2)

CouponUsed decimal	OrderCount decimal	DaySinceLastOr... decimal	CashbackAmount decimal	MembershipLevel text
4.0	6.0	7.0	139.19	Not Member
0.0	1.0	0.0	120.86000000000001	Not Member
2.0	2.0	0.0	122.93	Not Member
0.0	1.0	2.0	126.83000000000001	Not Member
1.0	1.0	1.0	122.93	Not Member
9.0	15.0	8.0	295.45	Not Member
0.0	1.0	0.0	153.81	Bronze
2.0	2.0	2.0	134.41	Not Member

Figure 3: Data Profiling in Talend Data Prep (Part 3)

By clicking through the charts, values, and advanced charts, I can get a better understanding of the data.

The Goals I was hoping to assess using Talend Data Prep were:

1. Formatting Issues
2. Missing Values
3. Using Advanced Chart to get a better idea of whether there is need for me to further explore in depth for outliers using SAS.
4. Suitability of data for Modeling

Issues Identified

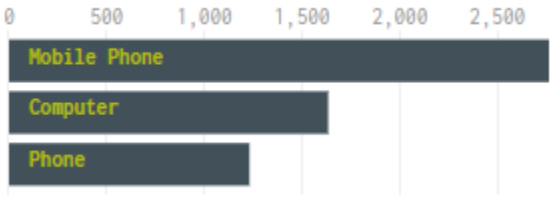


Figure 7: Chart of PreferredLoginDevice

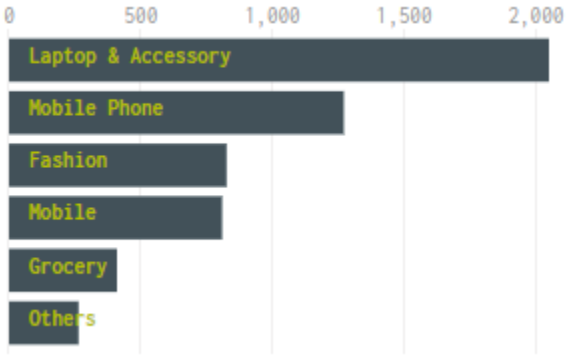


Figure 8: Chart of PreferedOrderCat

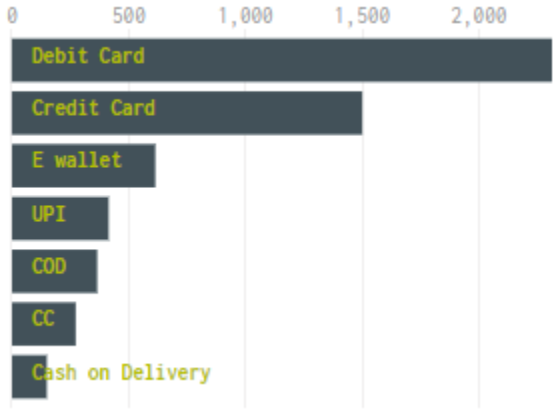


Figure 9: Chart of PreferredPaymentMode

CashbackAmount ☰ decimal
159.93
120.9
120.28
134.07
129.6
139.19
120.86000000000001
122.93
126.83000000000001
122.93
295.45
153.81
134.41
133.88
196.19
120.72999999999999

Figure 10: Rounding Inconsistencies in CashbackAmount.

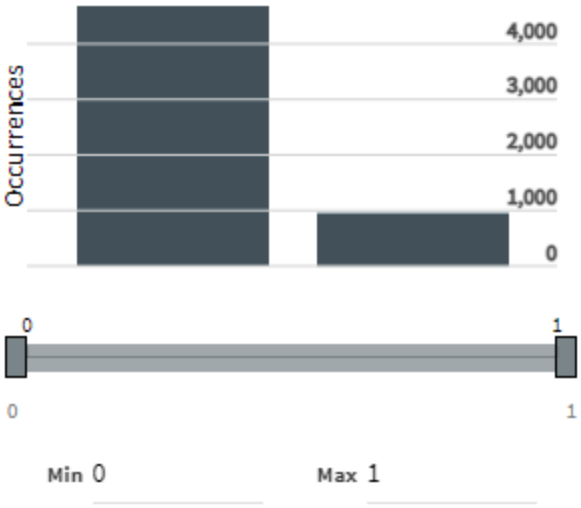
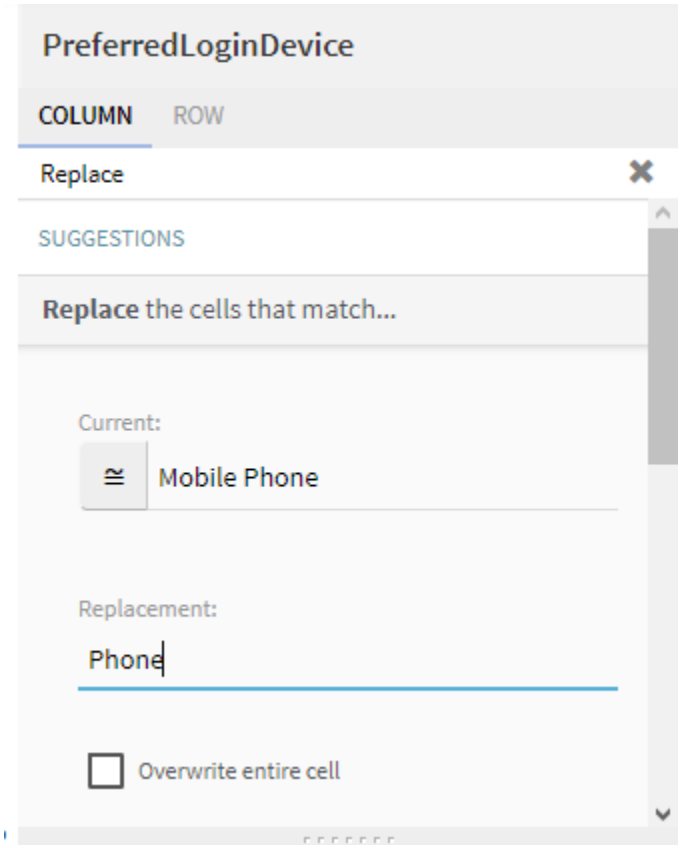


Figure 11: Imbalance Issue in Churn Variable

Steps to correct Issues in Talend Data Prep:



Future 12: Using Replace Function to Correct

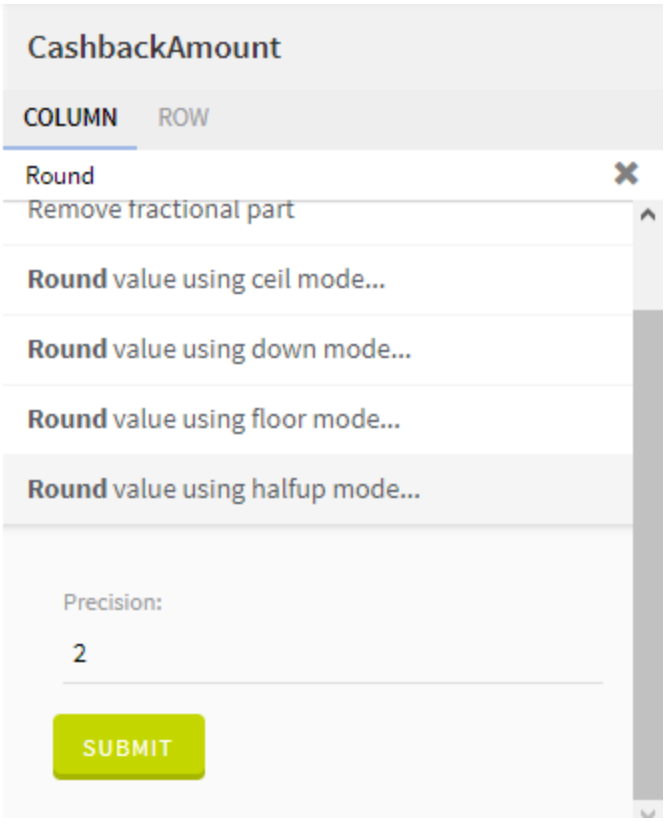


Figure 13: Rounding Values to Precision 2

Modify with Talend Data Prep

With the beginning exploration done, some of the issues identified will then be addressed in Talend Data Prep.

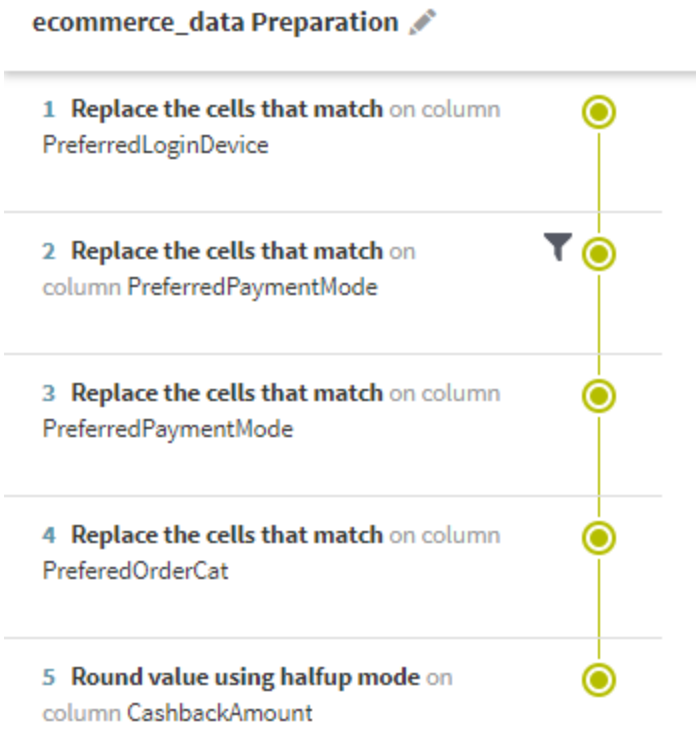


Figure 14: Correction of Issues in Talend Data Prep

As shown in figure 12, the naming inconsistent issues were first corrected using the replacement function. The text “Mobile Phone” was replaced with “Phone” in the PreferredLoginDevice variable.“CC” was replaced with “Credit Card” in the PreferredPaymentMode variable. “COD” was replaced with “Cash on Delivery” in the PreferredPaymentMode variable.“Mobile Phone” was replaced with “Mobile” in PreferedOrderCat variable. Lastly, the rounding format inconsistency as shown in figure 13 was corrected using Round value using half up mode with precision being 2.