How to Go from Zero to Hero in Two Hours with SAP Analytics Cloud

ANA264

EXERCISE 3:

How to perform data preparation with SAP Analytics Cloud.

Las Vegas:
Ian McAlpine / Matthias Badaire

Bangalore:
Jayanth Sunder Kaskyap / Narashimman K S / Samarth Narayan

Barcelona:
Stephane Perdigeon / Sven Frick

SAP TECHED

TABLE OF CONTENTS

INTRODUCTION

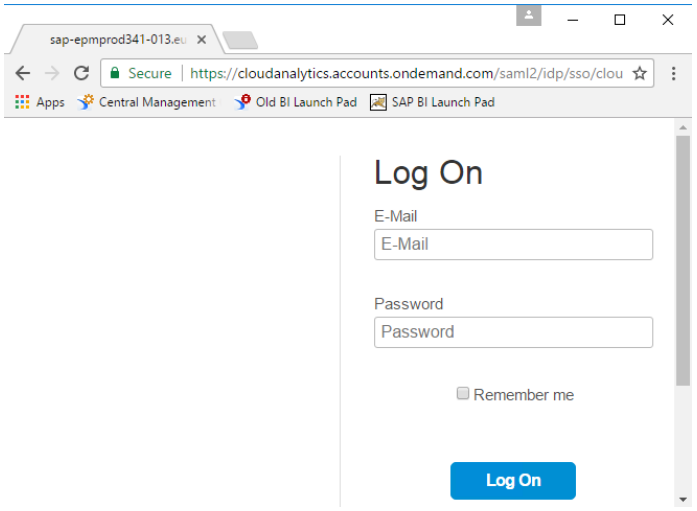
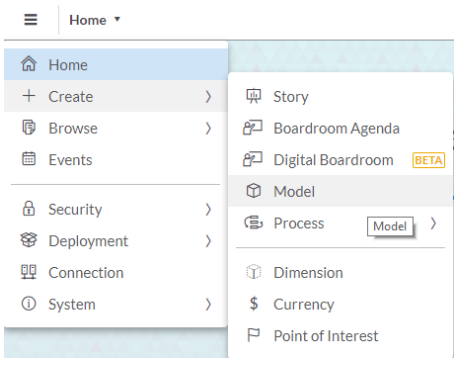
In this exercise, we will perform the following steps:

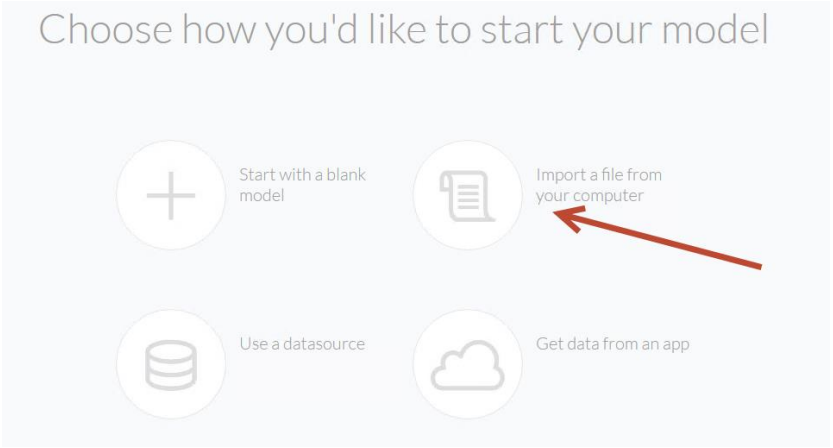
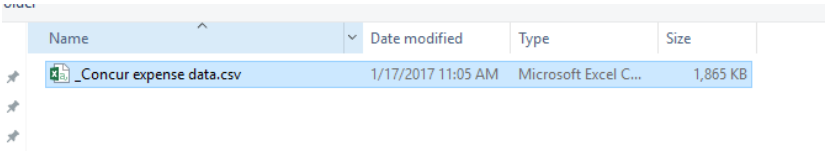
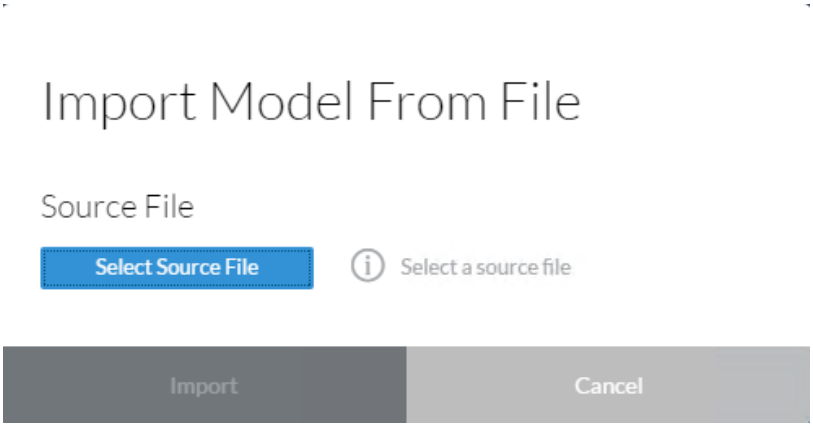
- Create a model based on a CSV file
- Perform some data manipulations and data cleansing on a sample of the data
- Validate data quality on the full dataset and build a model
- Create a story

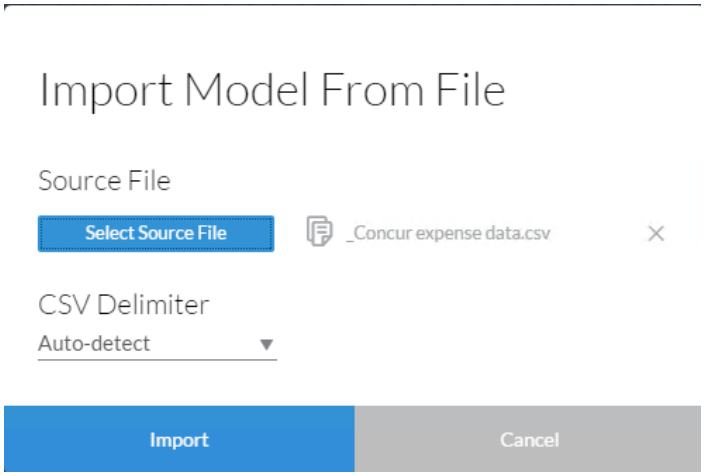
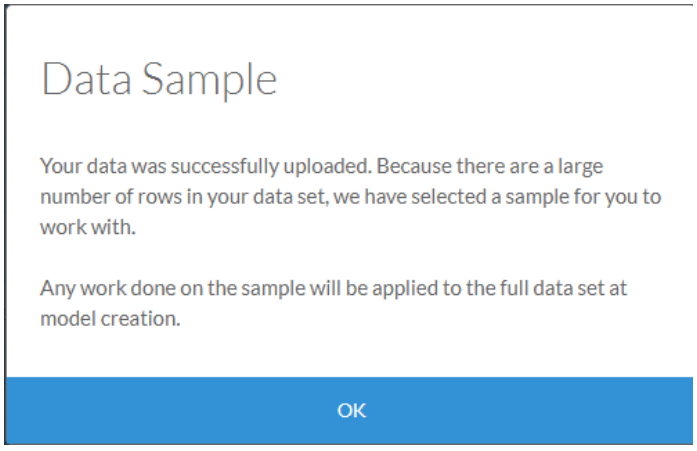
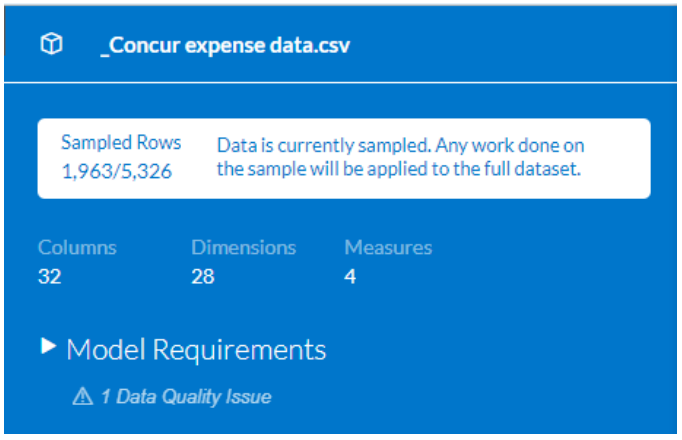
By the end of the exercise you will experience the new SAP Analytics Cloud modeler.

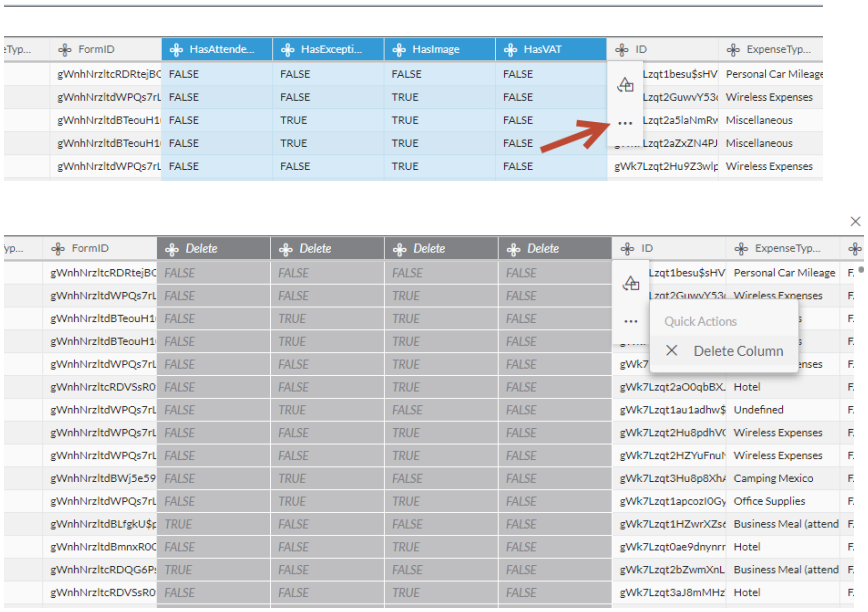
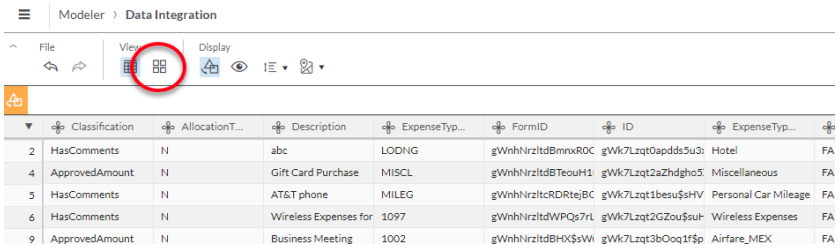
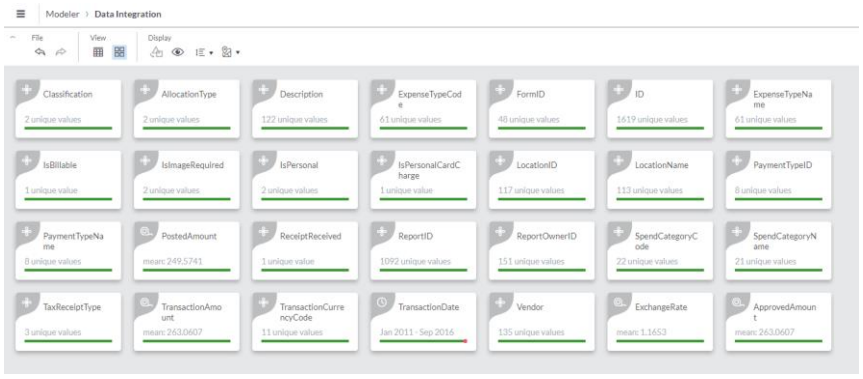
CREATE A MODEL BASED ON A CSV FILE

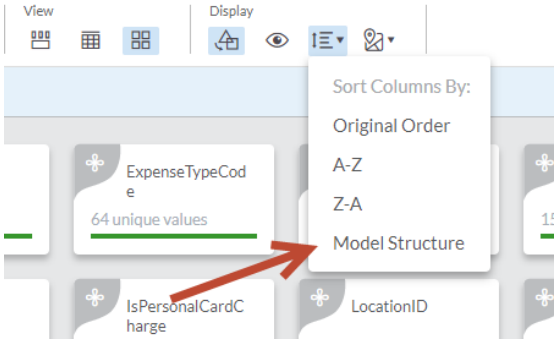
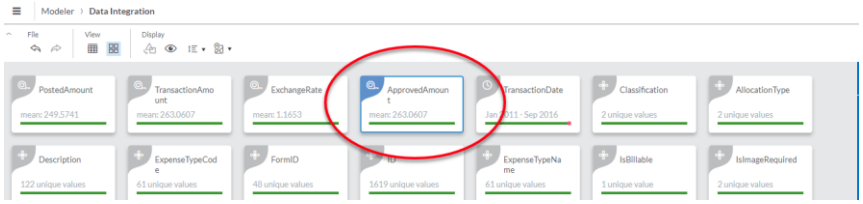
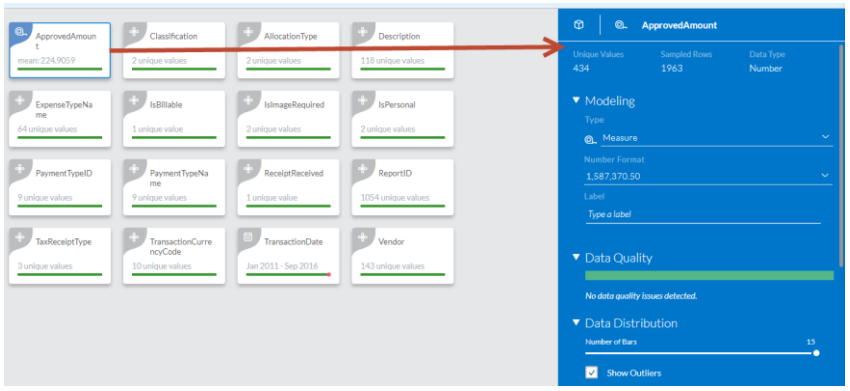
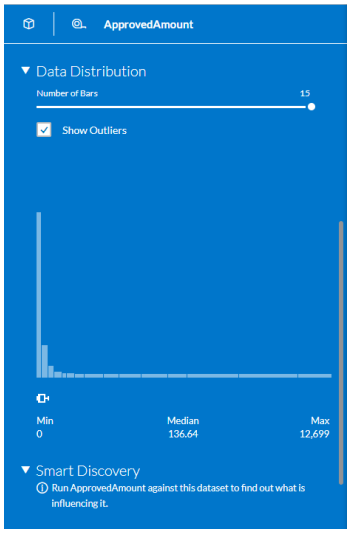
We start by creating a new model.

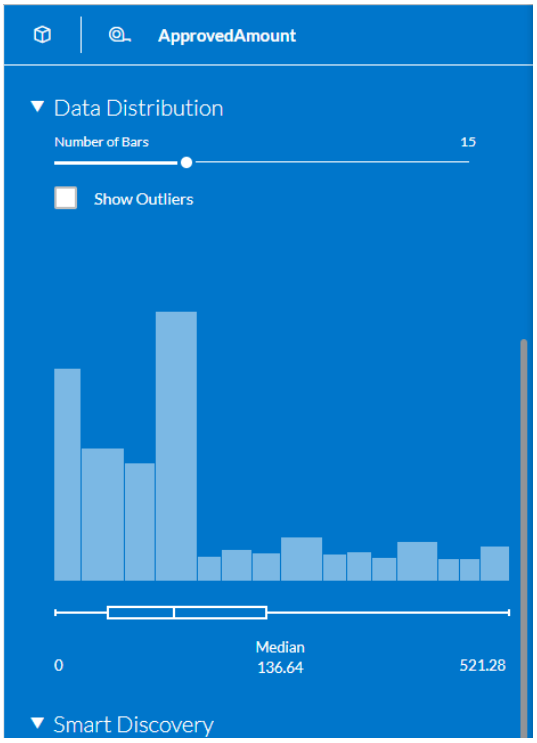
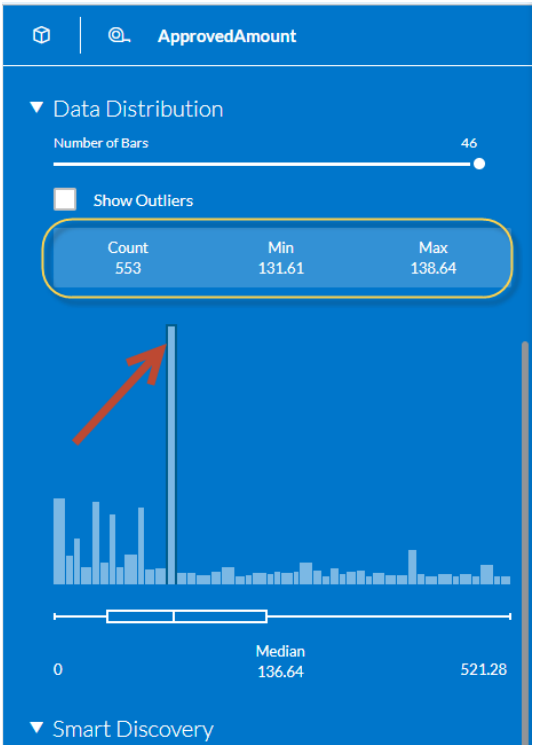
Explanation	Screenshot
Start Chrome and enter the following URL to logon to SAP Analytics Cloud: *https://sap-epmprod341-013.eu1.sapbusinessobjects.cloud* Log into the system using your credentials: **teched17ANA264+XXX@gmail.com** …where XXX is the number assigned to you (for example if you have 089, the email address will be: teched17ANA264+089@gmail.com) Password: **Password1**	
Click on *Hamburger* → *Create* → *Model*.	

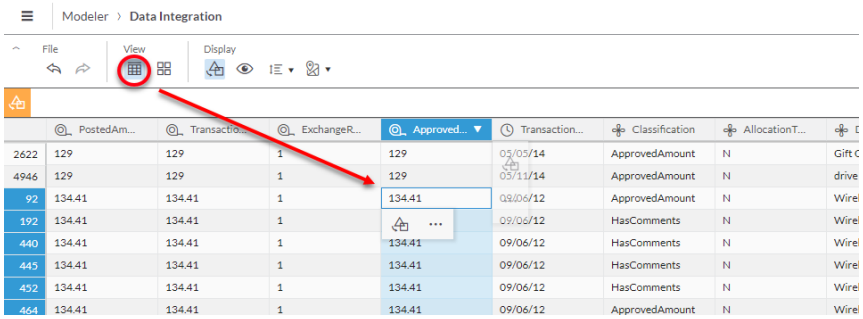
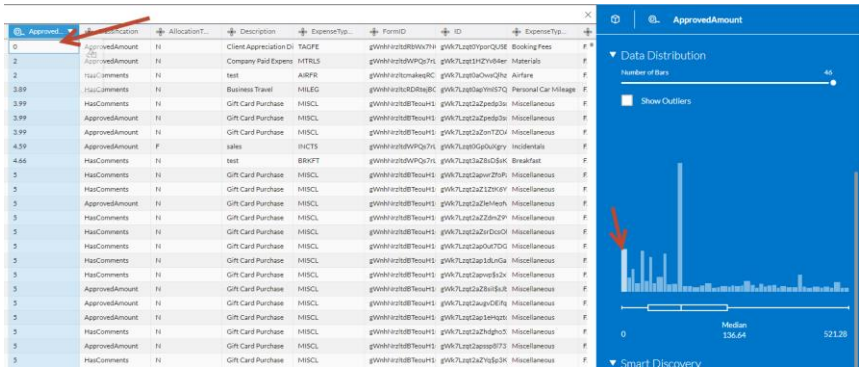
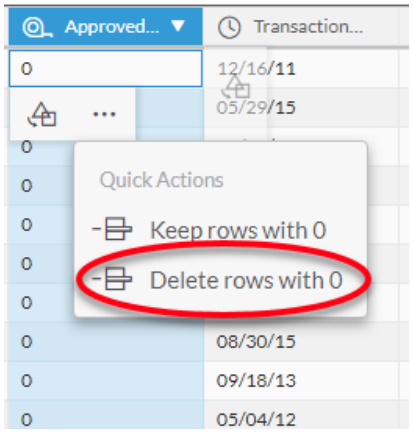
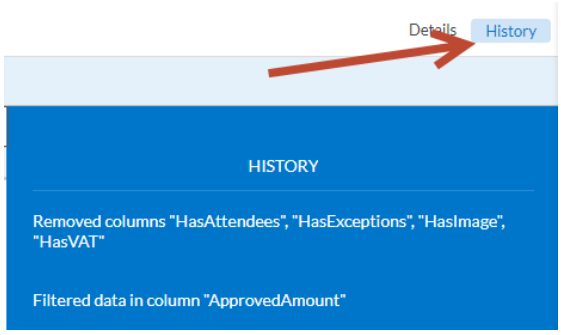
Explanation	Screenshot
Click on *Import a file from your computer*.	
Go to the following network folder to get the required CSV file: *\\students.fair.sap.corp\Students hare \ANA264* Find the file: *_Concur expense data_TechEd_2017.csv*. Copy this file to the Windows Desktop.	
Click Select Source File. Choose the *_Concur expense data_TechEd_2017.csv* file from the Desktop.	

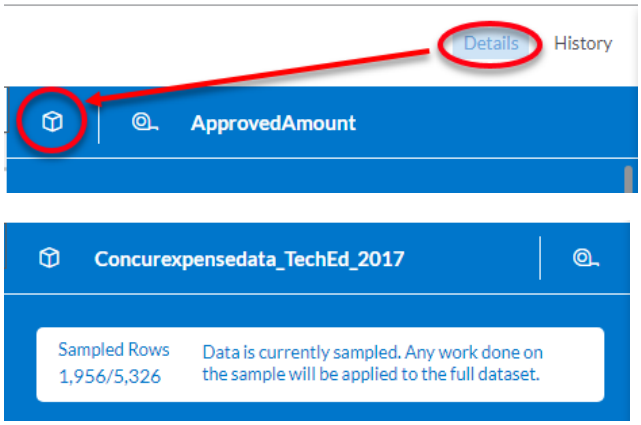
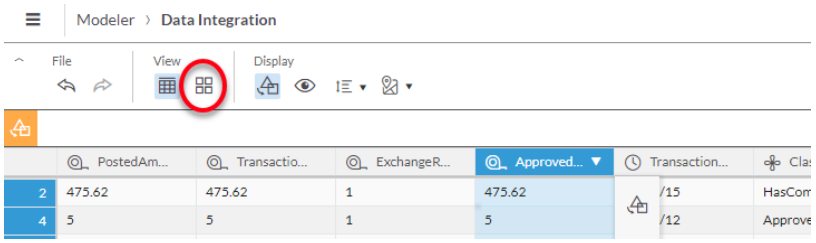
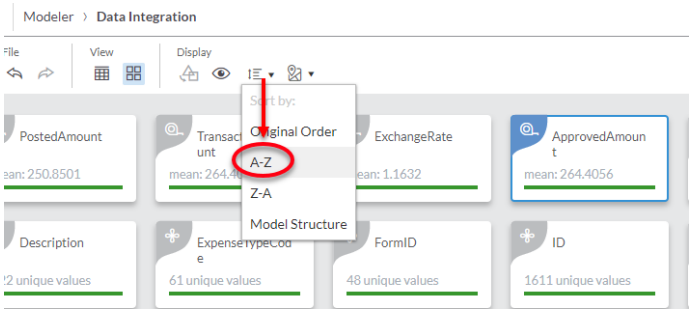
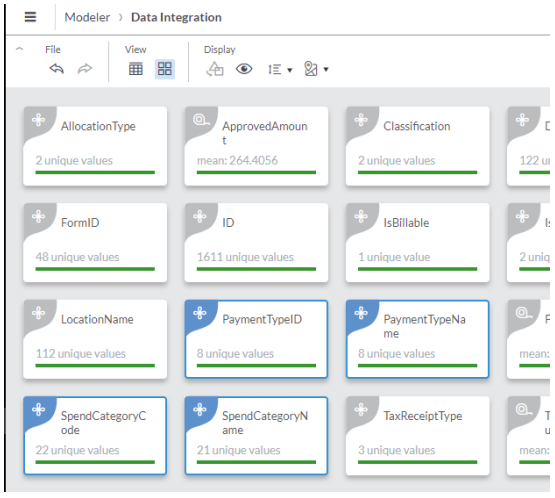
Explanation	Screenshot
Leave the default options and click *Import*. The system will start to load the data from the CSV file.	
Once the data is acquired we can view it. However, in this case because there is a lot of data we will be shown a sample. Click *OK*.	
The right-side panel of the modeler provides details about the dataset. We can see a sample size of 1,963 rows were loaded into SAP Analytics Cloud while the total row count in the CSV file is 5,326. Notice the total number of columns in the dataset.	

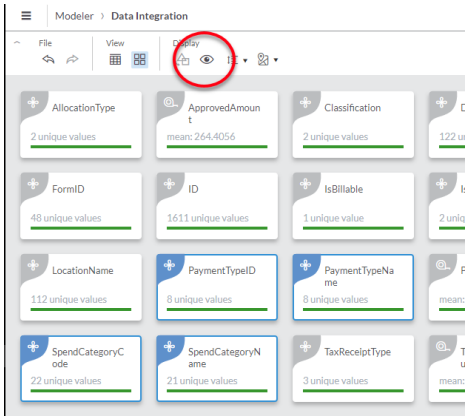
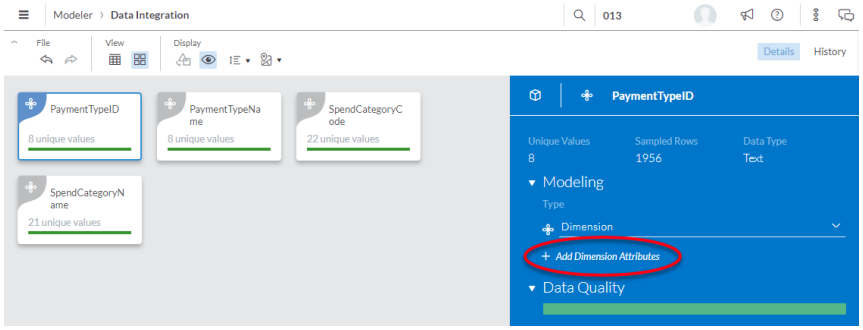
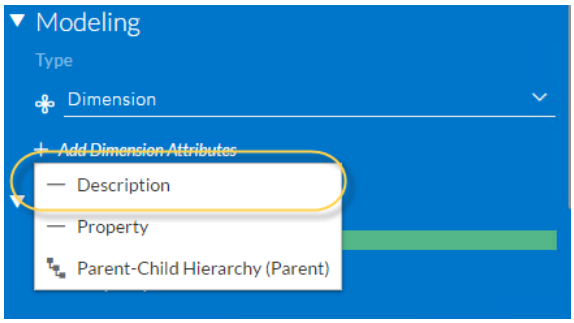
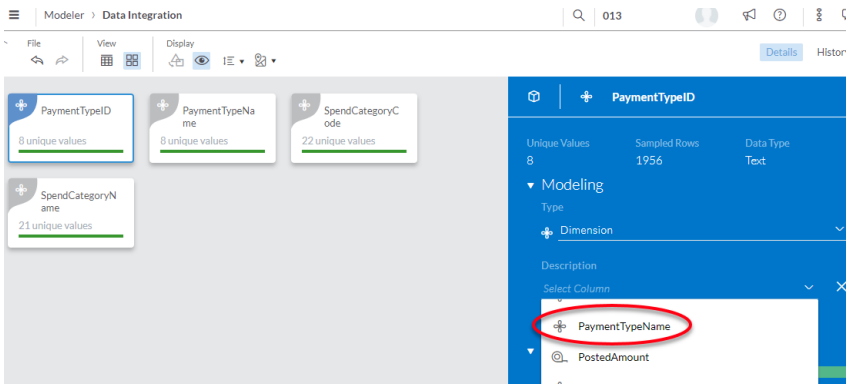
Explanation	Screenshot
We will start to do some modifications on the dataset. We will delete several Boolean columns that are not useful for analysis. Shift select *HasAttendees* through to *HasVAT*. Click the Quick Action button (the … icon) and choose the option *Delete Column*. Also note the preview that shows the columns to be deleted.	
The large number of columns makes it difficult to work in Table view with scrolling left and right. Instead we will work in card view. Click on the  icon to switch to card view.	
Note that different types of data are summarized in each card and each card also has a different icon for dates, categorical data and measures. • All the data types are inferred • Measures show their mean • Categorical values show number of unique values • Dates show range	

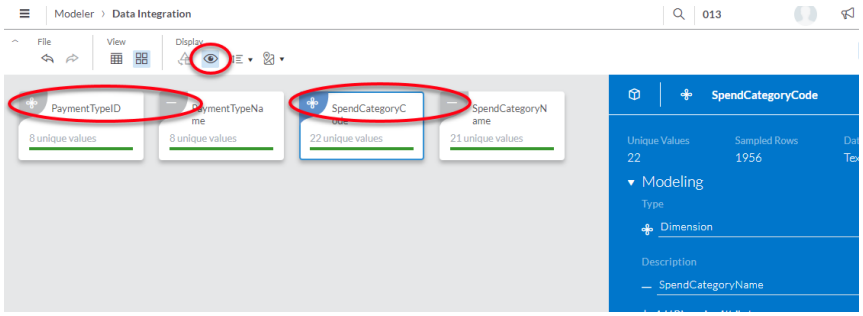
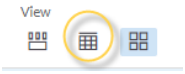
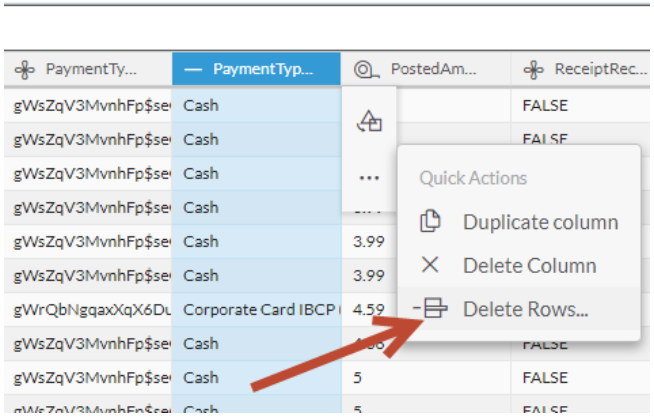
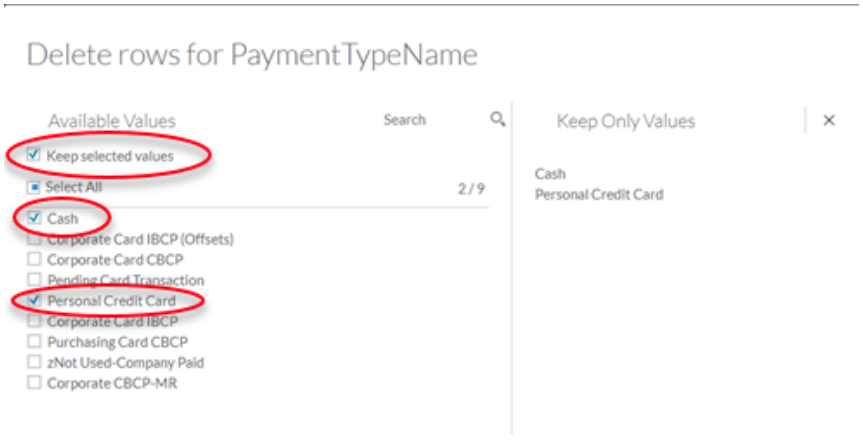
Explanation	Screenshot
In the toolbar click on the *Sort* icon  Choose *Model Structure* to display all the measures at the top of the card view.	
Click on the *ApprovedAmount* measure (4th card).	
Notice how the right panel changes to reflect the information about the selected card. Scroll down in the right panel to get additional information on Data Distribution.	
Hover the pointer over the bars in the Data Distribution chart and notice that there are some large expenses (ones with small count values) that are skewing the results.	

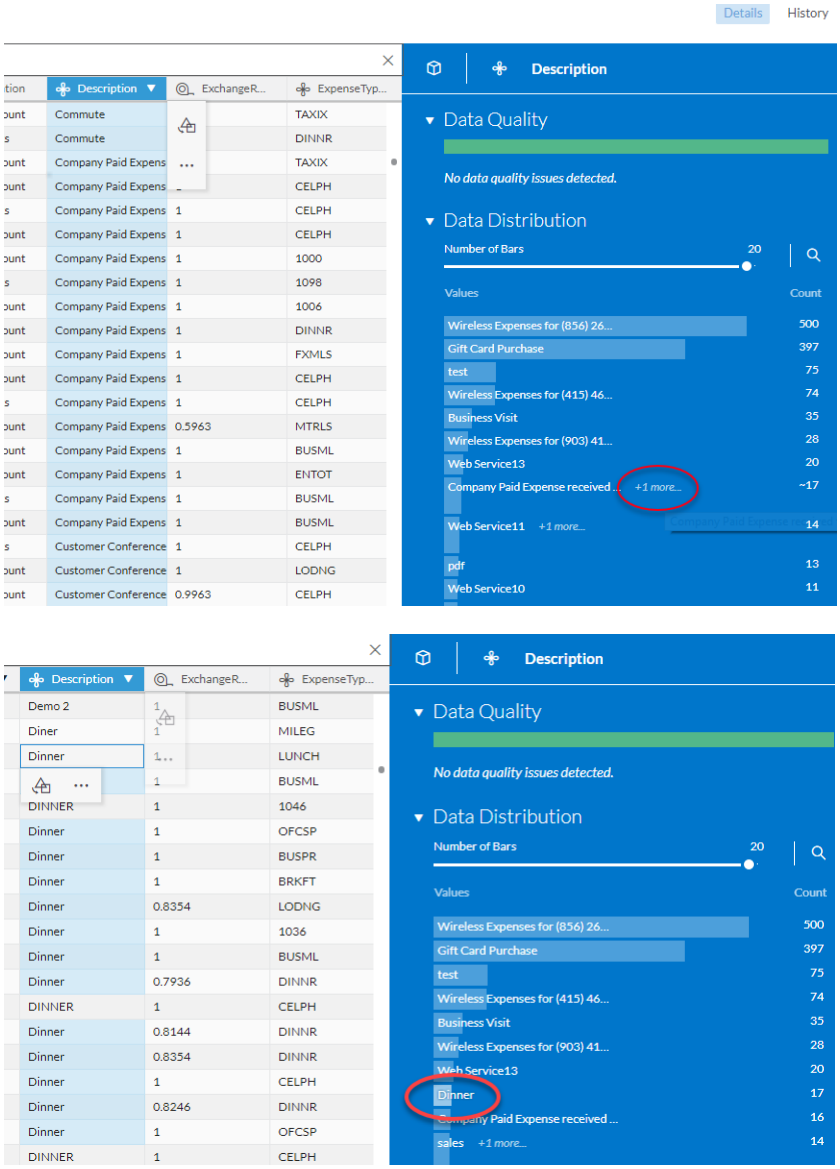
Explanation	Screenshot
Uncheck *Show Outliers* and notice how the focus changes to show the bulk of the data.	
Increase the number of clusters from 14 to the maximum by dragging the bar to the right. Aside: If you cannot use the mouse to drag the bar, use the right arrow key on your keyboard. Notice the interesting peak (highest bar) around $130 and select it.	

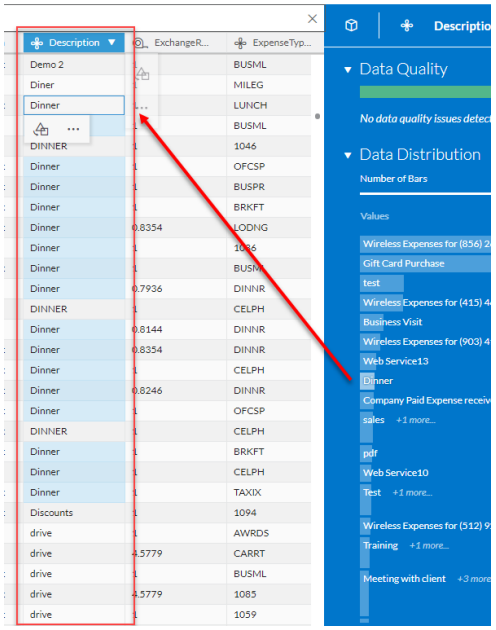
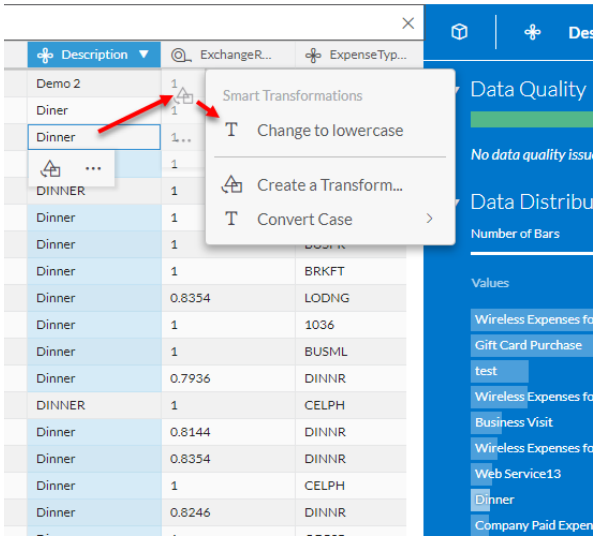
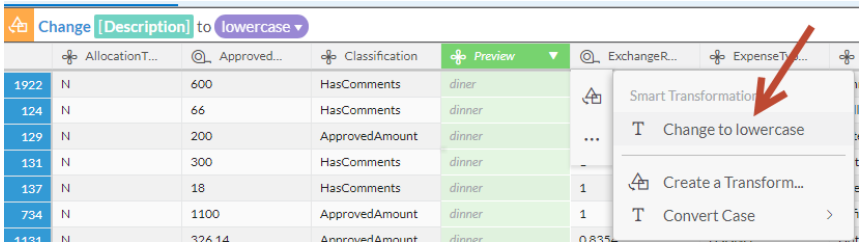
Explanation	Screenshot
Click the Table View icon in the menu bar to view the data.  Note what seems to be a set of duplicate entries which may require further investigation.	
From the <u>right panel</u>, select the first bar showing the lowest values. The focus of the *ApprovedAmount* column will shift to show the 0 values.	
Select one of the *0* values. Click on the *Quick Action* icon (…) and select *Delete rows with 0*. CAUTION: Make sure to click on the *Quick Action* icon (…) icon <u>below</u> the cell (which acts on the selected cells) and not the one on the right (which acts on the whole column).	
Click on *History* to see the history of the transformations for this column.	

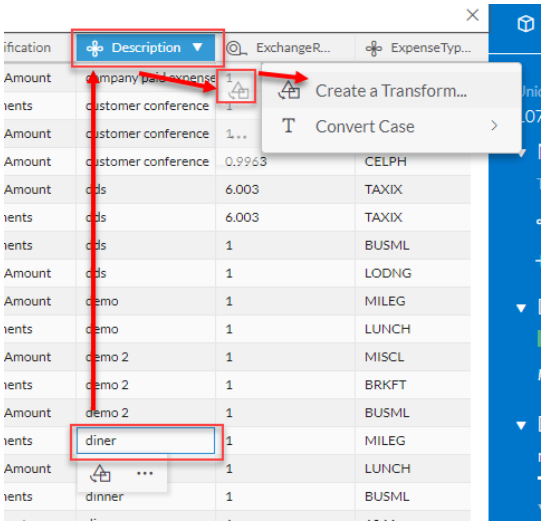
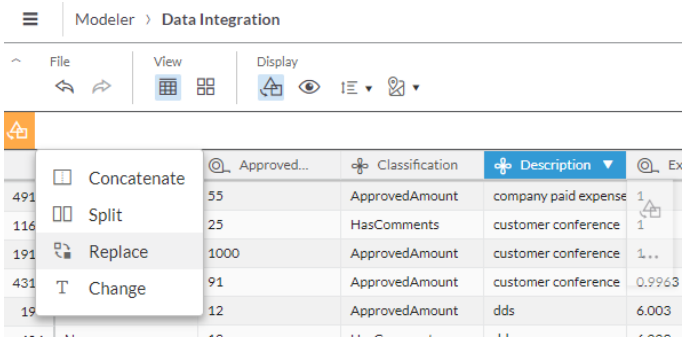
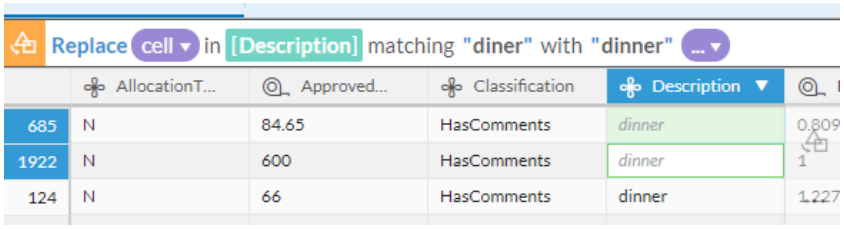
Explanation	Screenshot
Click on *Details*. Click on the *Model Details* icon to view the history on the whole dataset. 	
In the menu bar, click on the *Card View* icon to switch back to the card view.	
Click on the *Sort* icon and choose *A-Z*. This will allow us to easily see related concepts that start with the same name.	
Click on *PaymentTypeID*. Press and hold the <Ctrl> key and click on the *PaymentTypeName*, *SpendCategoryCode* and *SpendCategoryName* cards. Four cards should now be selected.	

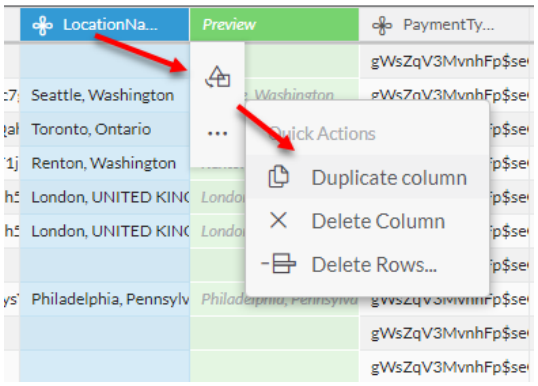
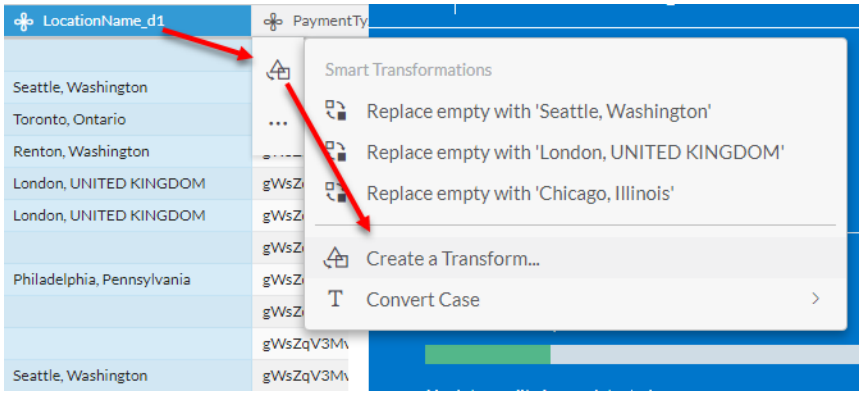
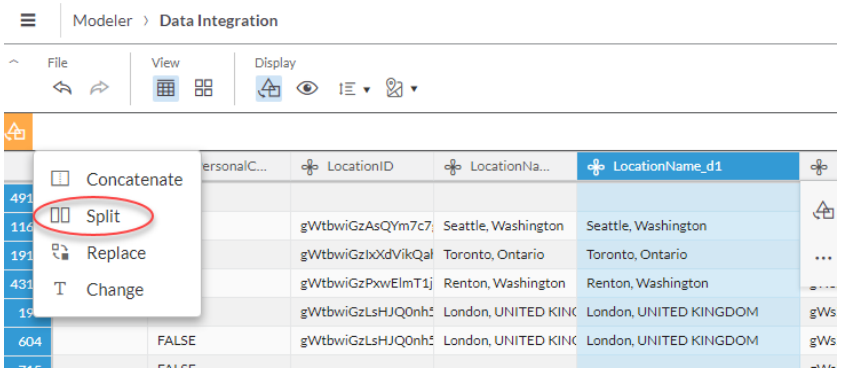
Explanation	Screenshot
Click on the *Focus* icon to show only the four selected cards.  This allows you to only focus on the dimensions you need to work on (especially useful when you have many columns in your model).	
We are going to link the “name” dimension as an attribute of the “ID” dimension. Select the *PaymentTypeID* card. In the right hand panel, expand the Modeling section and click *Add Dimension Attributes*.	
Select *Description*.	
Select column *PaymentTypeName*.	

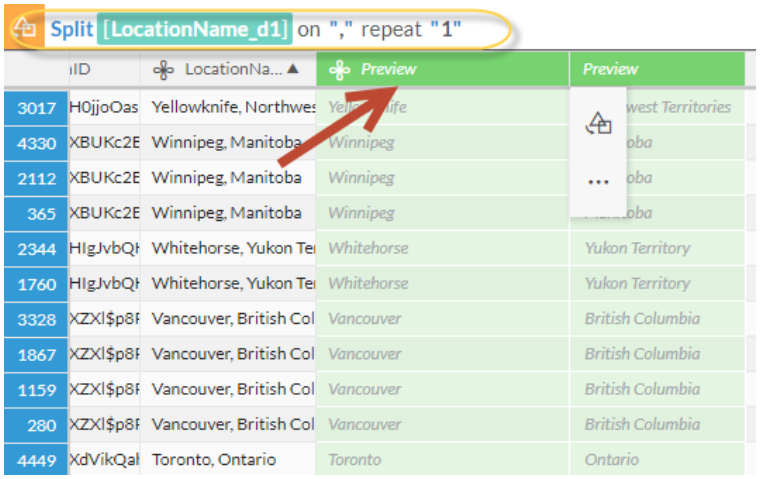
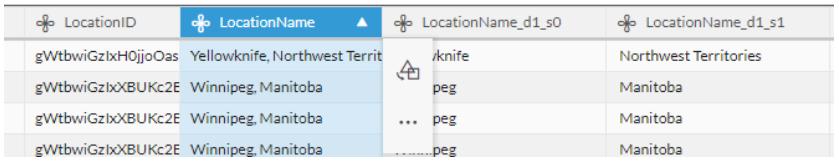
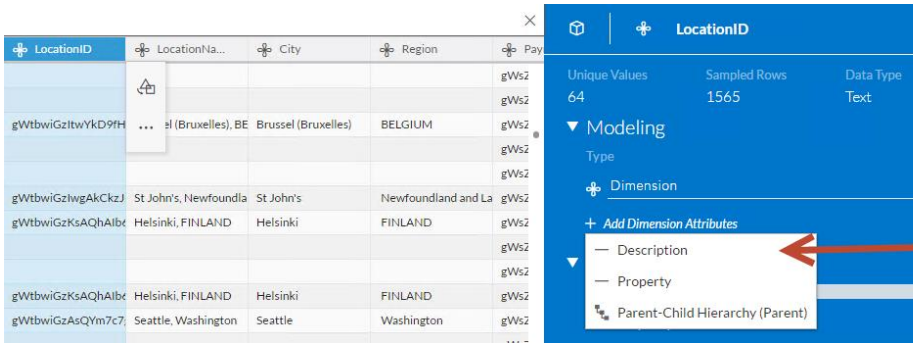
Explanation	Screenshot
Repeat the process for *SpendCategoryCode*. Click on the *SpendCategoryCode* card and set *SpendCategoryName* as its description. Note the change in the card icons as the columns are now linked. Click the *Focus* icon again to go back to the full dataset.	
Click the *Table View* icon.  Select the *PaymentTypeName* column. Click the *Quick Action* icon (…) and choose *Delete Rows...*	
Ensure the option *Keep selected values* is checked. Check *Cash* and *Personal Credit Card* then click *OK*. The above action will delete all rows where the PaymentTypeName is not "Cash" or “Personal Credit Card”.	

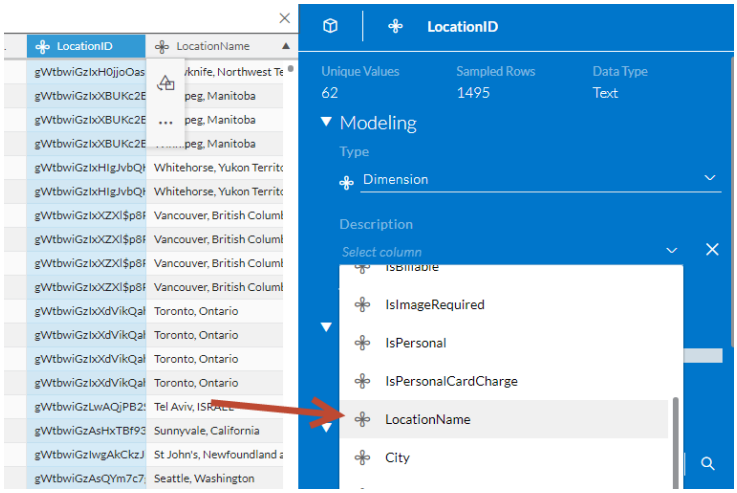
Explanation	Screenshot
Select the *Description* column. From the right panel, click on *Dinner*. Note: You might have to click on *+1 more…* on some lines to find it.	

Explanation	Screenshot
The focus of the *Description* column is automatically shifted to the selected values. Notice that some entries for *Dinner* have a different case, plus there are also spelling mistakes.	
Click on the *Transform* icon  in the Quick Action menu and choose *Change to lowercase*. Notice the preview in the column before you click on it. Caution: the *Transformation* icon is available at both the cell and column level. Make sure to select the column level icon as indicated in the screenshot. All the values for Dinner will now change to be *dinner*.	 

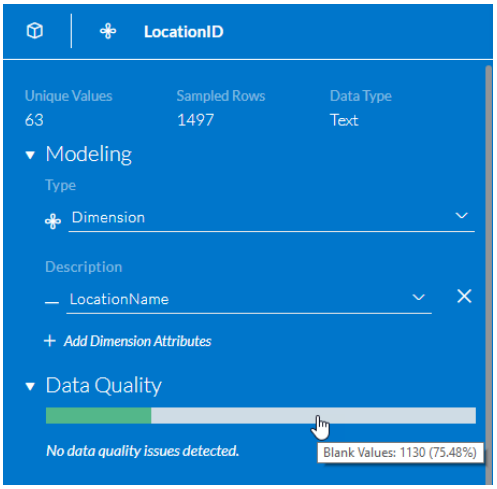
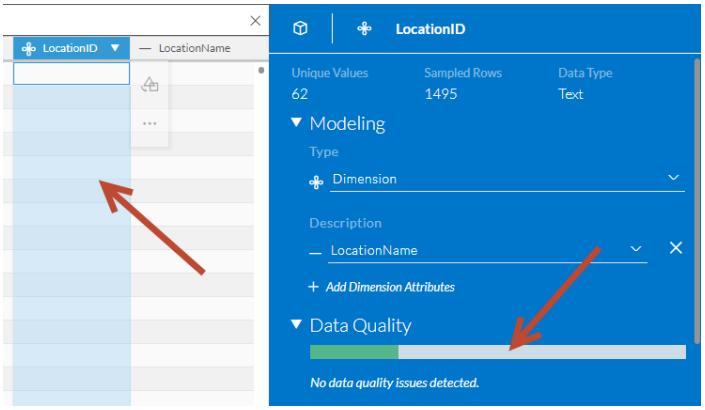
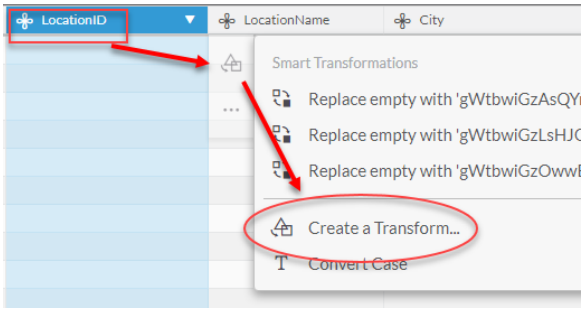
Explanation	Screenshot
We will now fix the incorrect spelling for *dinner*. Select one of the cells with incorrect spelling for *dinner*. Click the Transform icon  at the column level. Choose *Create a Transform…*	
For the transformation, choose *Replace*.	
Put the cursor inside "*value*" and type *dinner*. Notice the preview on the selected cell. Press the <Enter> key on your keyboard when done. All entries for *dinner* should now have correct case and spelling.	

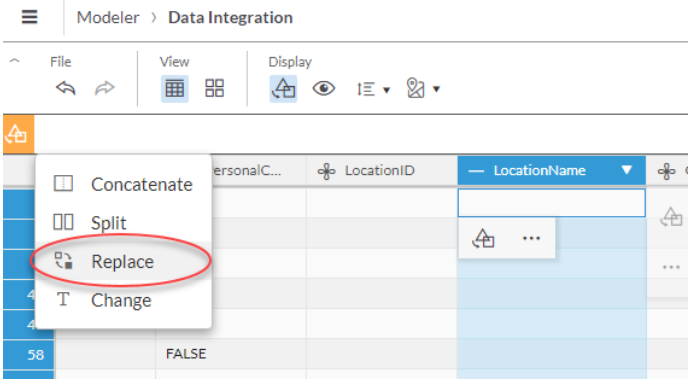
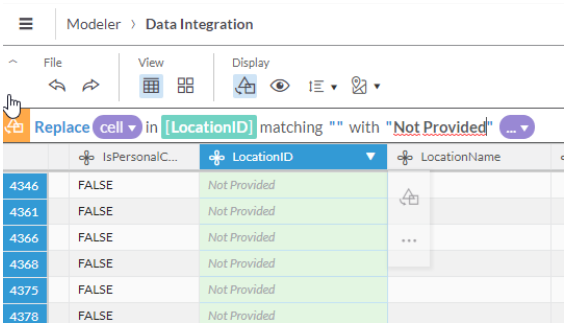
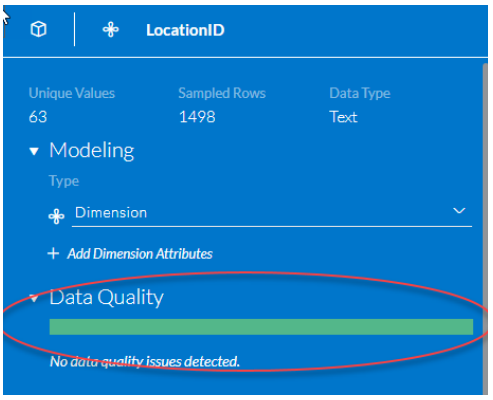
Explanation	Screenshot
Now we are going to split the *LocationName* column into two columns to get *City* and *State* columns. First, we duplicate the *LocationName* column. Select the *LocationName* column. Click on the *Quick Access* icon (…) and select *Duplicate column*. Notice a preview column appears.	
Select the newly created column *LocationName_d1*. Notice that it has both the City and State/Country information. Click the *Transform* icon and select *Create a Transform…*	
Select the *Split* transformation.	

Explanation	Screenshot
Place the cursor on "*delimiter*" and press the <,> (comma) key. Notice again the two new preview columns. The split will act on the *LocationName_d1* column and split it into two columns based on the "," delimiter. Press the <Enter> key on your keyboard.	
Two new columns are created; *LocationName_d1_s0* and *LocationName_d1_s1*.	
To rename a column, double click on the column header and enter the new name. Rename *LocationName_d1_s0* to **City**. Rename *LocationName_d1_s1* to **Region**.	
Similar to before, we will make *LocationName* an attribute of *LocationID*. Select the *LocationID* column. Under the Modeling section, click on *Add Dimension Attributes*. From the drop-down list choose *Description*.	

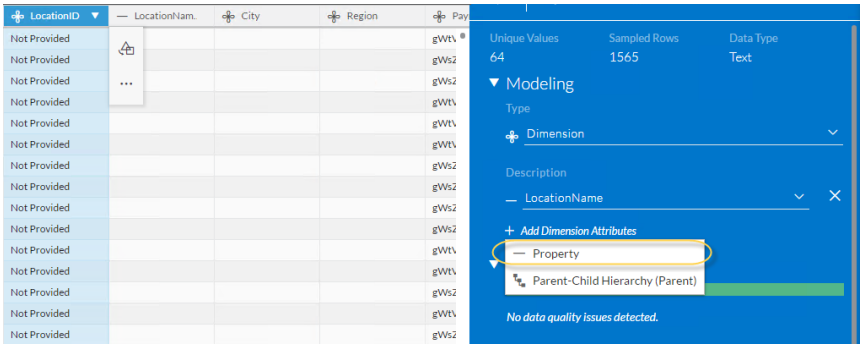
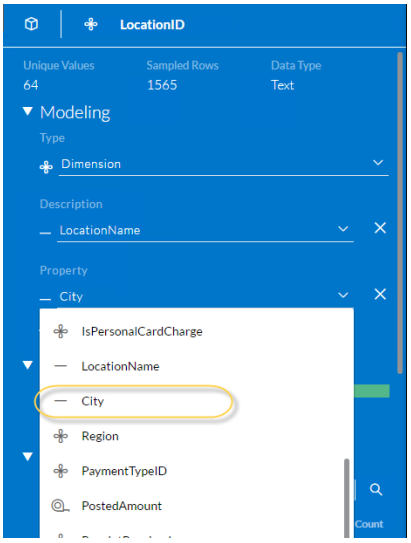
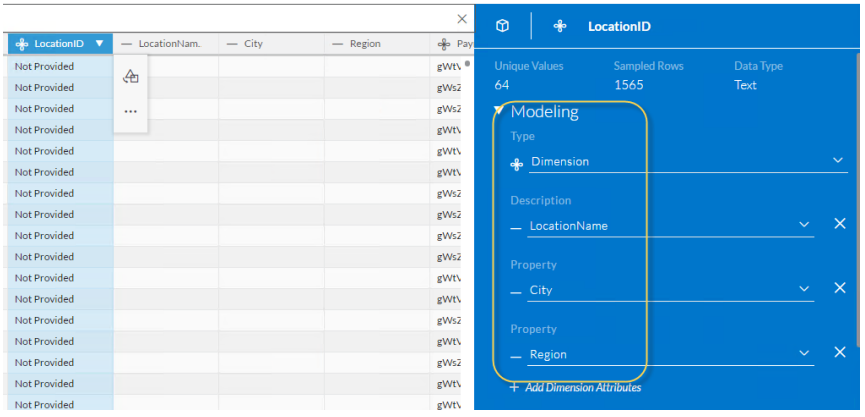
Explanation	Screenshot
Select the *LocationName* column as the description.	

CHECK DATA QUALITY

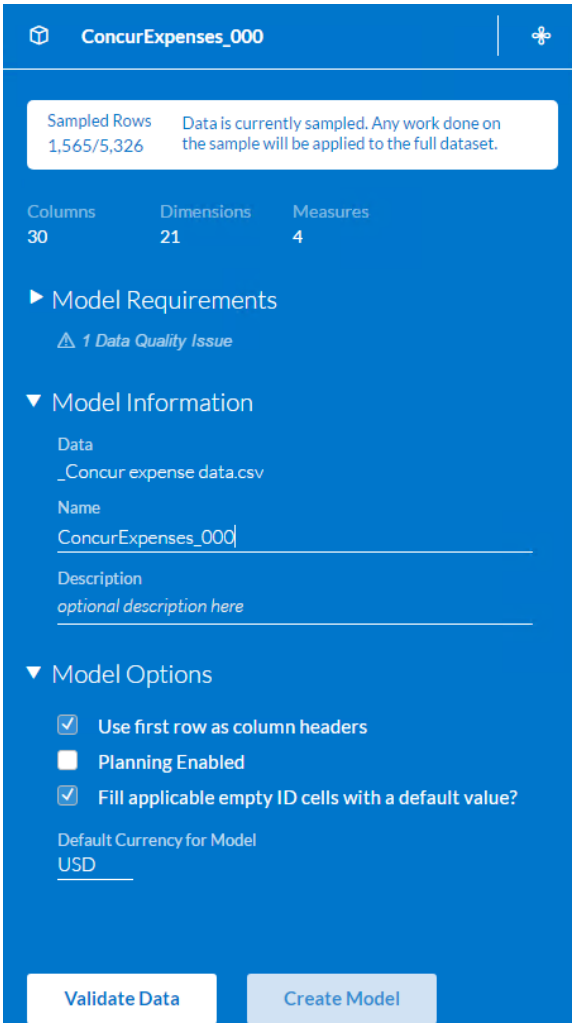
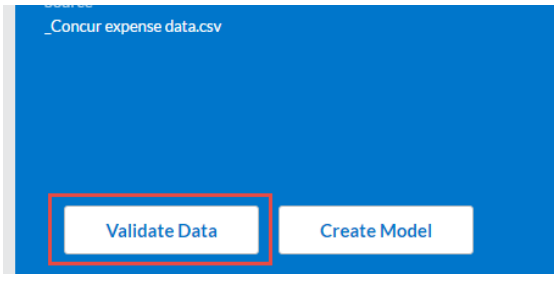
Explanation	Screenshot
Select the *LocationID* column. In the right-hand panel, notice the Data Quality bar is not fully green. This is because of empty values in the *LocationId* column. Hover over the different sections of the bar to see the messages.	
In the Data Quality bar, click on the blank area. The *LocationID* column focus will shift to show the blank cells.	
Select one of the empty cells. Click on the *Transform* icon  and select *Create a Transform…*	

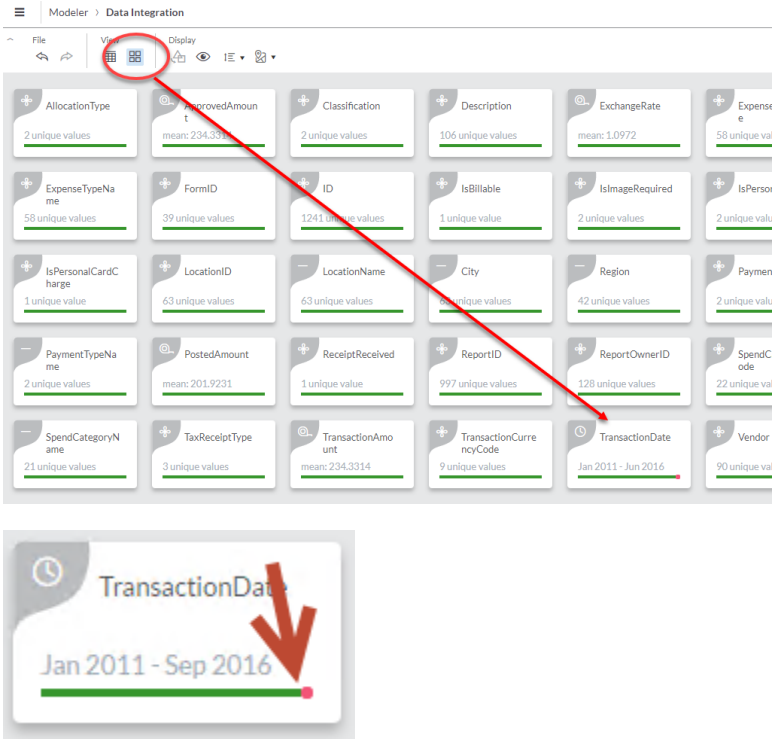
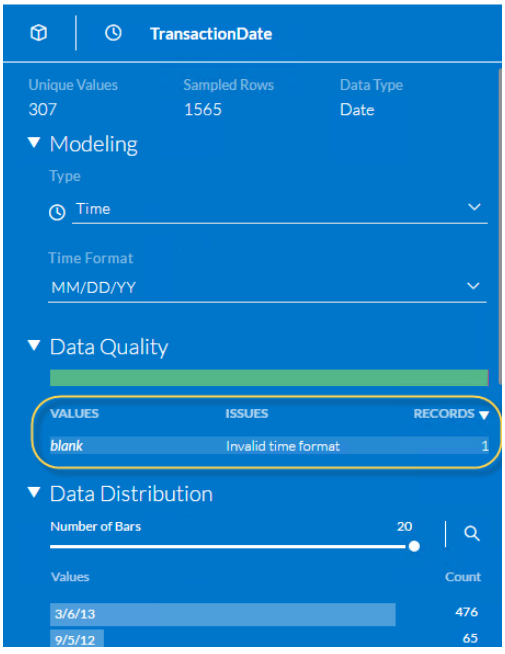
Explanation	Screenshot
Select the *Replace* transformation.	
In the transform rule, click on the first "*value*" so that it becomes empty. Click on the second "*value*" and enter **Not Provided**. Notice the preview column. Press <Enter> on your keyboard.	
The data quality issues should disappear.	

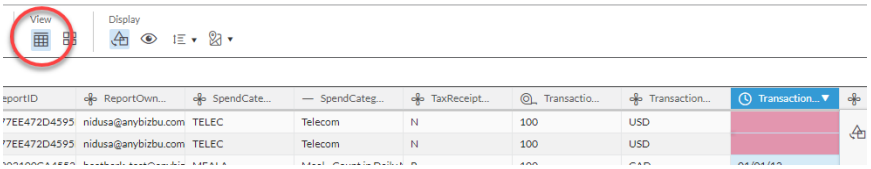
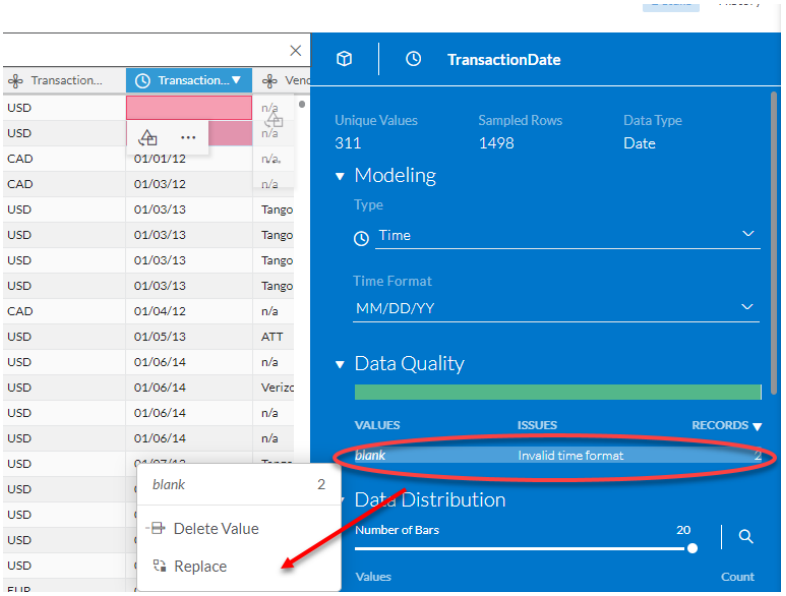
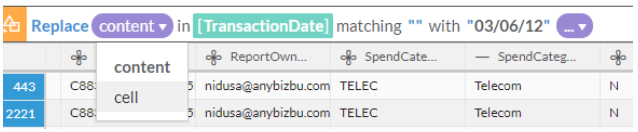
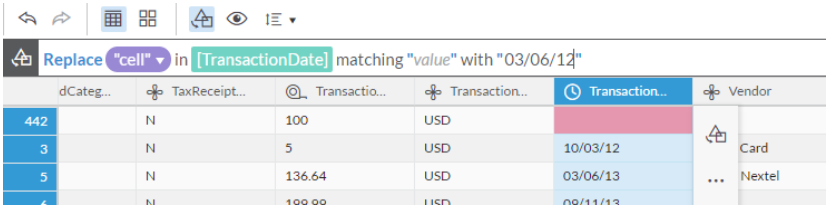
ADDING PROPERTIES TO A DIMENSION

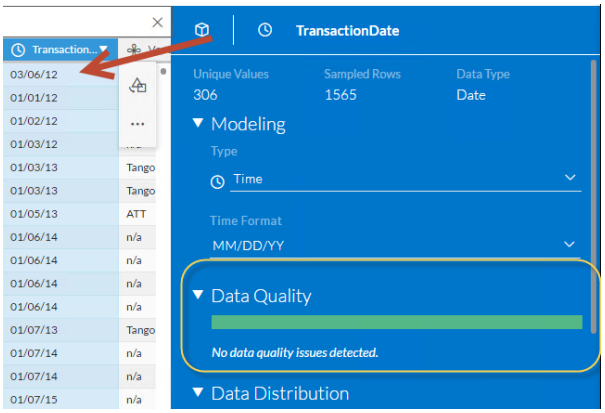
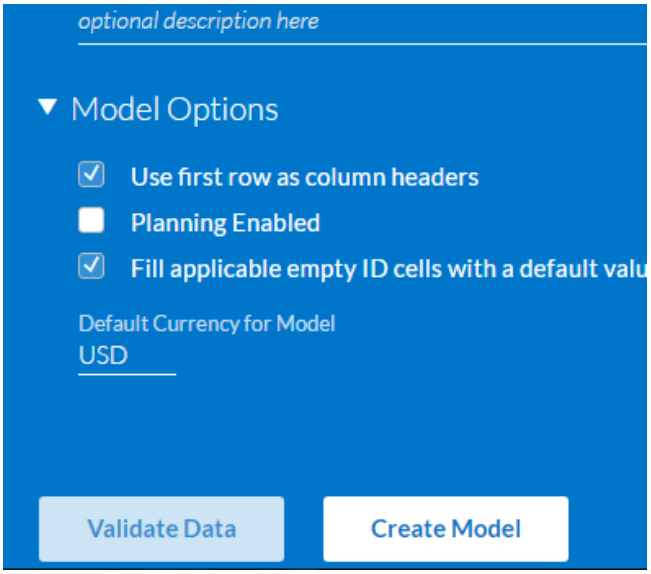
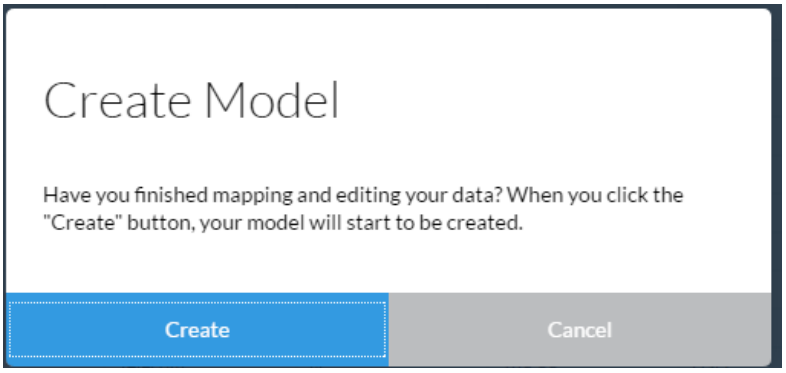
Explanation	Screenshot
Now we are going to set *City* and *Region* as properties of the *LocationID* dimension. Select the *LocationID* column. Under Modeling, click *Add Dimension Attributes*.	
Choose *Property*, then select *City*.	
Repeat the same operation with *Region*. The *LocationID* column settings should look like the screenshot.	

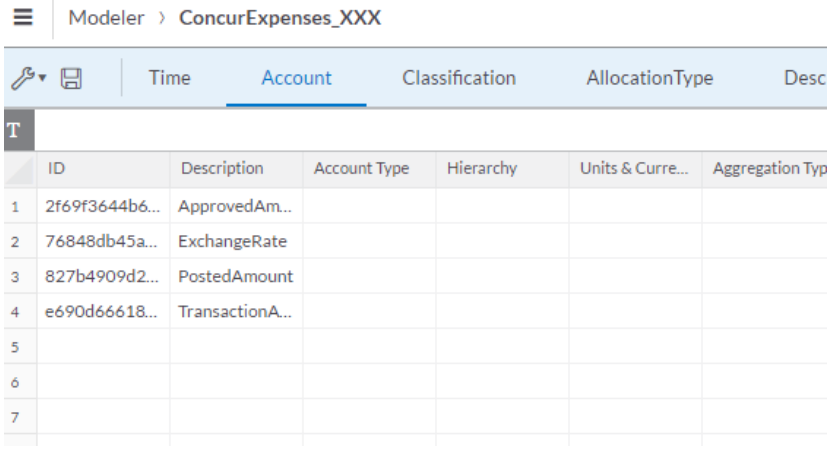
DATA VALIDATION

Explanation	Screenshot
Click the *Model Details* icon  to switch from the dimension view to see the model information. Expand the Model Information section and change the name from *Concurexpensedata_TechEd_2017* to *ConcurExpenses_XXX*, where XXX is the number assigned to you. This will ensure your model has a unique name and avoids naming conflicts with other users.	
At this point we can either create the model or go through validation first to check if additional data quality issues exist in the dataset. Remember, so far we have only been working on a sample of the dataset. The *Validate Data* options will confirm if additional quality issues exist in the entire dataset. Click on the *Validate Data* button.	

Explanation	Screenshot
The validation phase will add additional rows to the sample dataset which still have data quality issues. In this case one new row is added. Click on the *Card View* icon to determine that the quality issue is with the *TransactionDate* dimension.	
Click on the *TransactionDate* card, and you will get additional information.	

Explanation	Screenshot
Click on the *Table View* icon and notice one of the records is missing a value for *TransactionDate*.	
Click on the blank line in the Issues section. Actions are proposed. Apply a transformation of your choice to address this. Either choose *Delete Value* or apply a *Replace* transformation and enter any date value.	
We have chosen to replace the empty value with a random date. You might have to switch to *cell* if the default is *content*.	
Make sure the formula is as shown in the screenshot. Press <Enter> on your keyboard.	

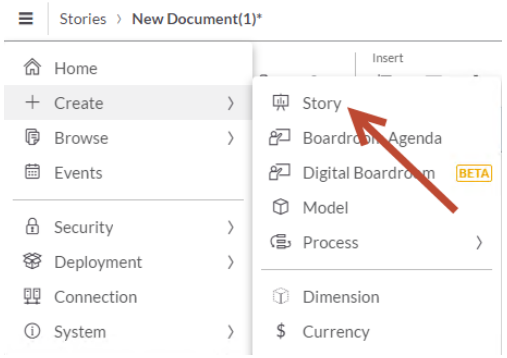
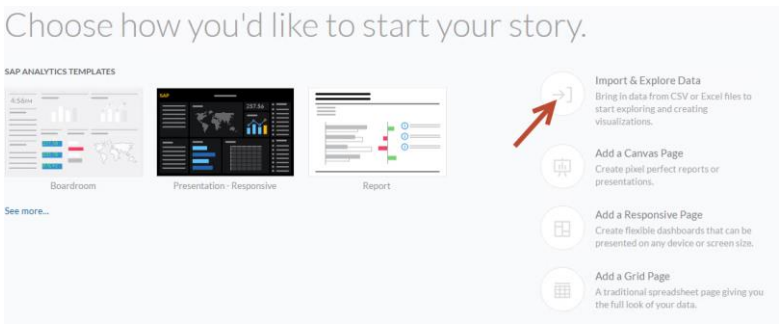
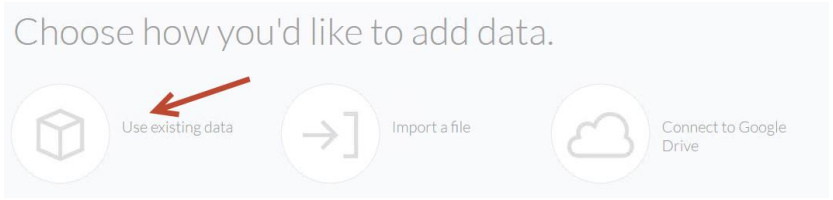
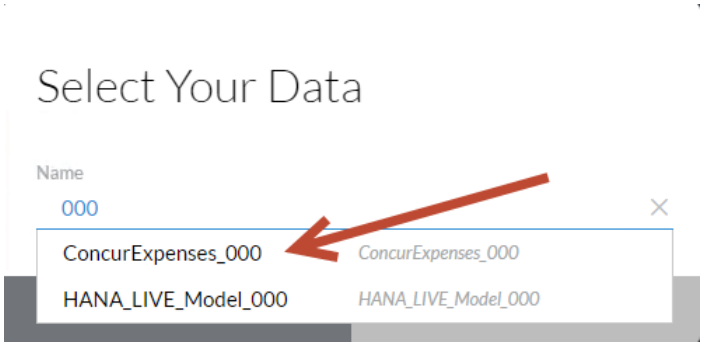
Explanation	Screenshot
All data quality issues should now be addressed.	
Click on the *Create Model* button to create the model. The Create Model process also goes through a validation phase. Since we have already validated the model on the full dataset, there should not be any data quality issues.	
Click *Create* on the warning message.	

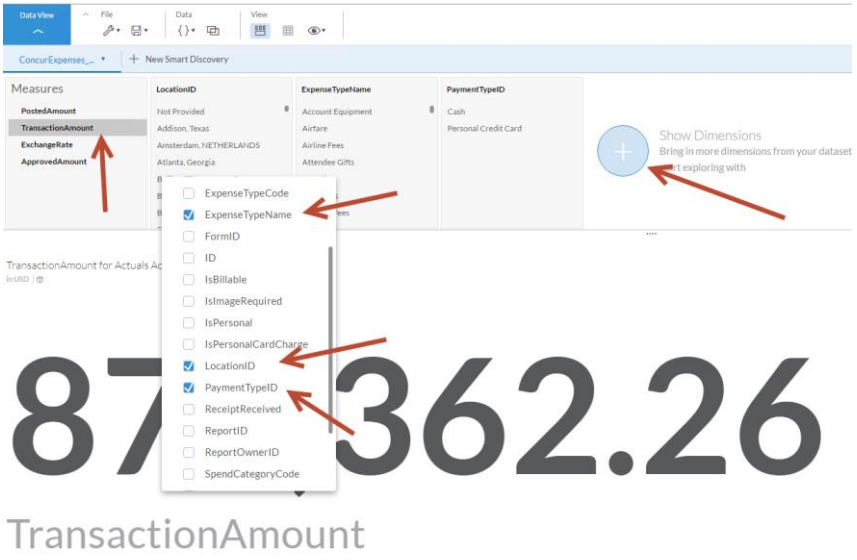
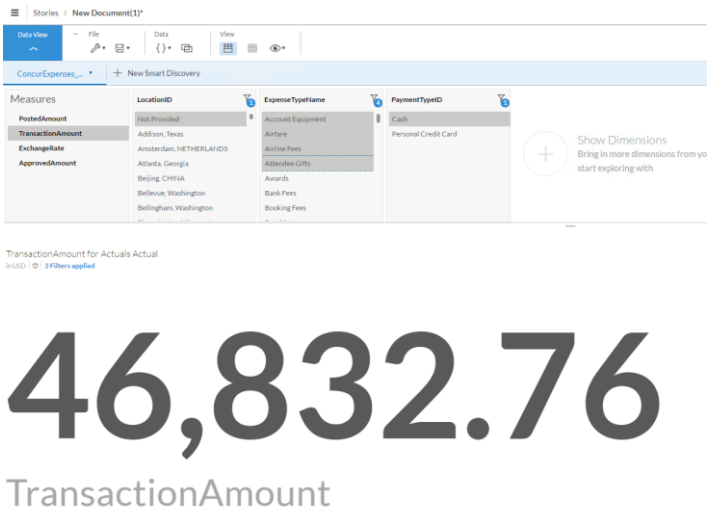
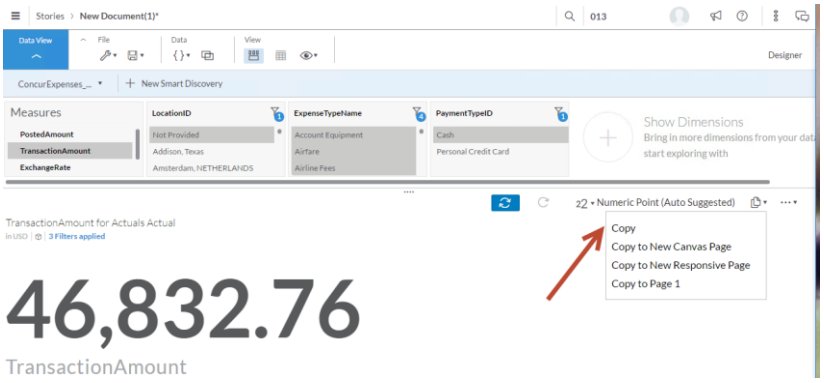
Explanation	Screenshot
The model is now created. Click the *Save* icon.	

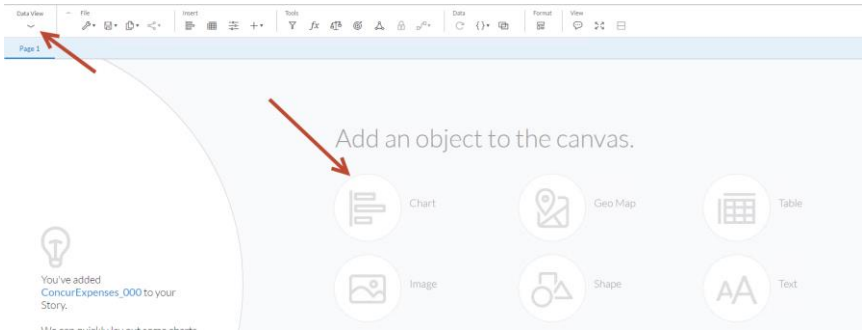
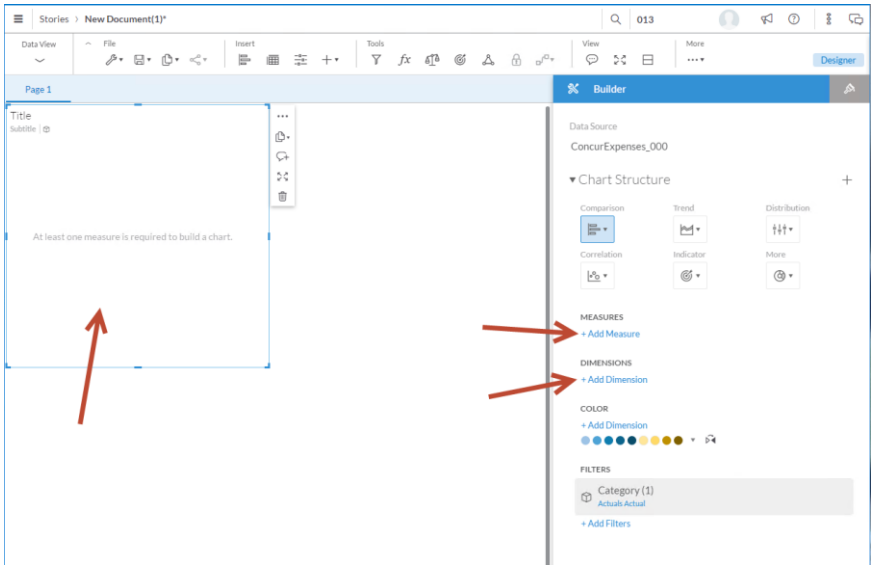
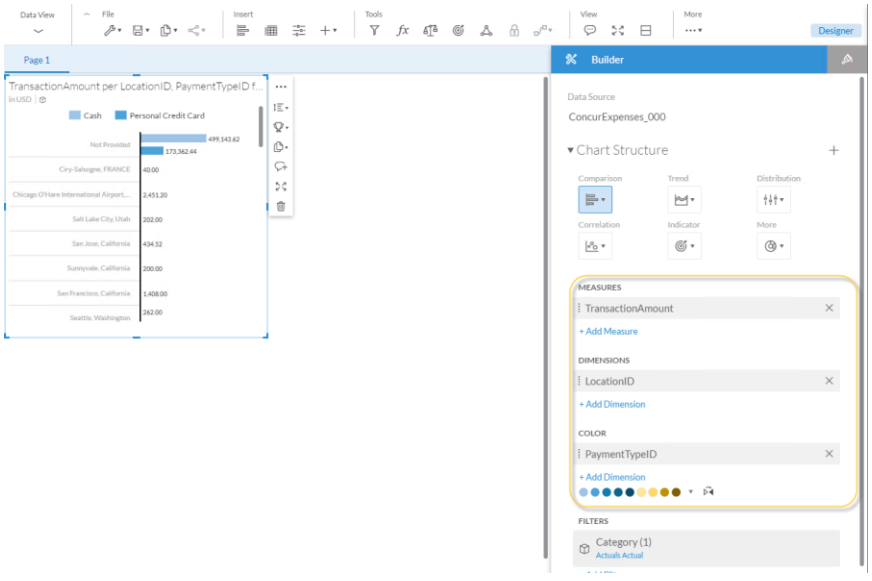
The model is now created.

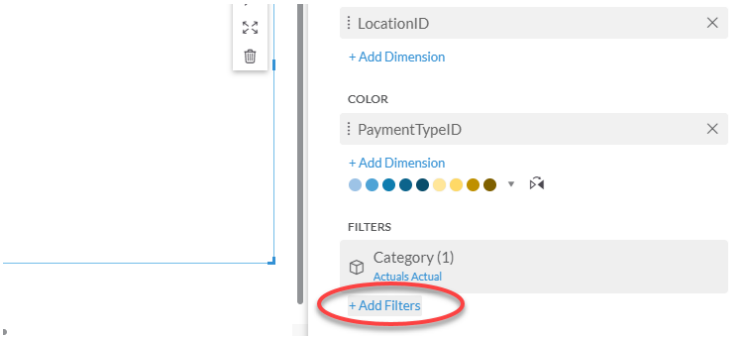
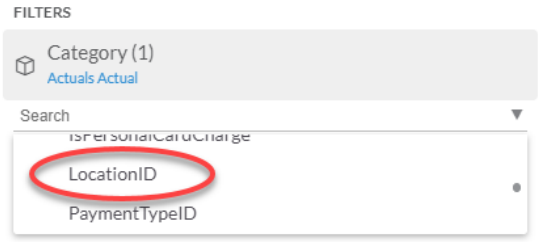
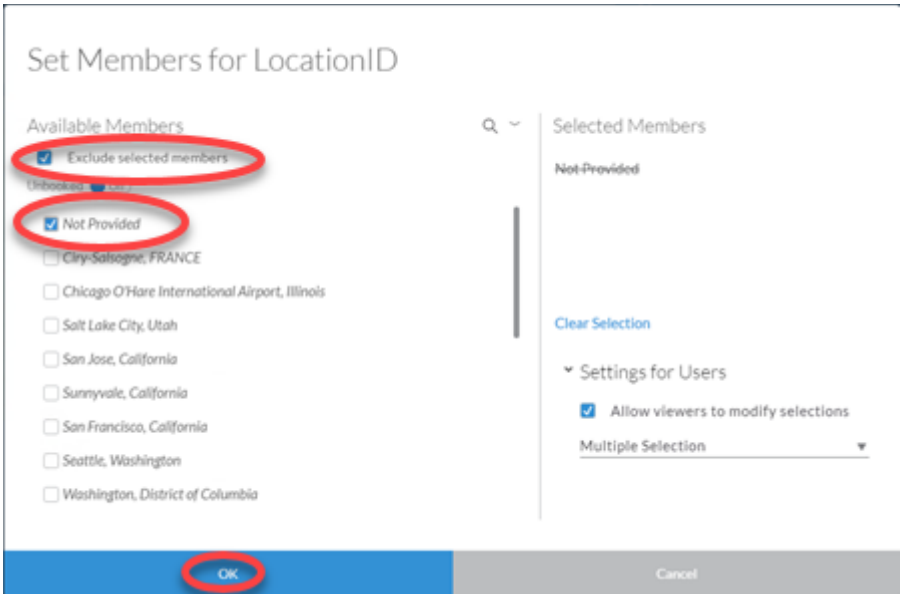
We will test the model by creating a simple story using the it.

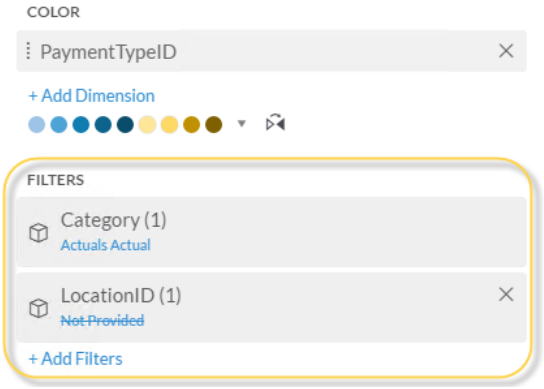
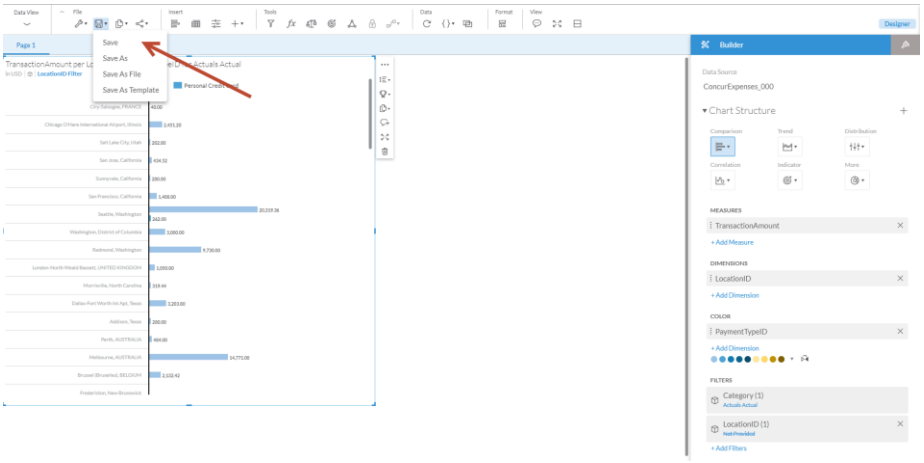
CREATE A STORY

Explanation	Screenshot
To create a story, click *Hamburger* → *Create* → *Story*.	
This time we select the option *Import & Explore Data*.	
Choose *Use existing data*.	
Select the model we just created, *ConcurExpenses_XXX*, where XXX is the number assigned to you.	

Explanation	Screenshot
Because we want to explore the data, the Data View appears. Here you can explore data by clicking on a Measure, for example *TransactionAmount* and then add the desired Dimensions. Click on *TransactionAmount*. Click on *Show Dimensions* and select the following dimensions: • *ExpenseTypeName* • *LocationID* • *PaymentTypeID*	
Now you can start exploring the data. Click on different members to filter the results.	
If you find something useful, you can copy the results to your canvas.	

Explanation	Screenshot
To close the Data View, click on *Data View* in the toolbar. Click on *Chart* to add a chart to the canvas.	
Select the chart and the Builder panel will appear on the right-hand side, where we can add Measures and Dimensions.	
Add the following to the chart: Measures: **TransactionAmount** Dimensions: **LocationID** Color: **PaymentTypeID**	

Explanation	Screenshot
The “Not Provided” member is skewing the results so we will add a filter to *LocationID* to exclude it. Click *Add Filters*.	
Select *LocationID*.	
Check *Exclude selected members*. Check *Not Provided*. Click *OK*.	

Explanation	Screenshot
You should have the following result.	
Click on the Save icon. Save your story as *DataPrep_XXX*, where XXX is the number assigned to you.	

www.sap.com/contactsap