Mark Friedman
University of Miami

Using Tableau Prep to Join Data Streams

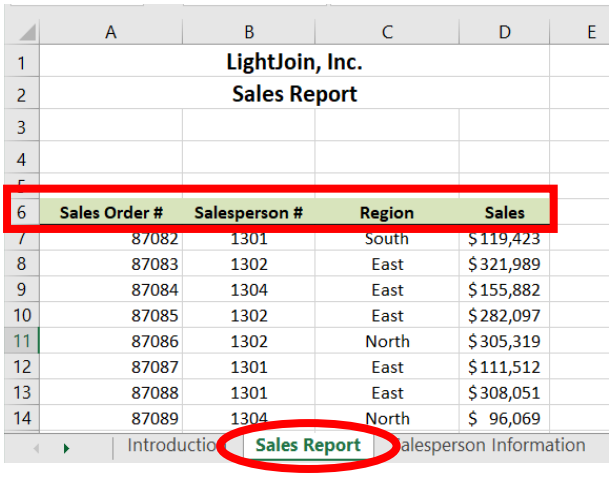
Tableau Prep is used by several Big Four accounting firms to automate the ETL process. In this exercise, you will:

Extract the "*Sales Report*", "Salesperson Information" and "Comm SalesPerson by Region" data from Excel.

Transform the data into a flat data set in which all the data for a specific sales order is on one row. Aggregate the commission in dollars for each salesperson by region.

Load the transformed data into Excel.

There are three streams of data for *LightJoin, Inc.* stored in an Excel workbook. The first stream, sales information is stored in a sheet titled "**Sales Report**". Notice that the column headings are in the sixth row.



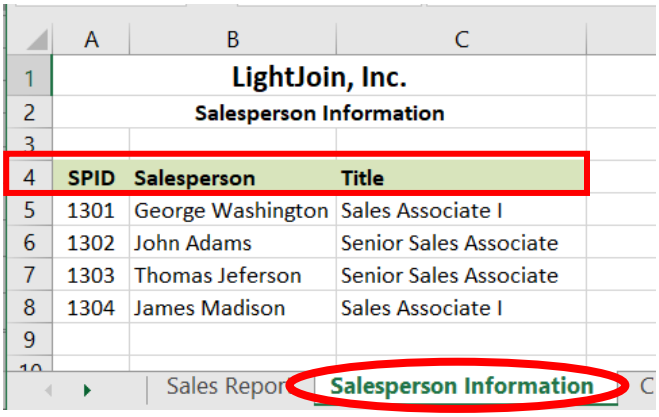
	A	B	C	D	E
1		LightJoin, Inc.			
2		Sales Report			
3					
4					
6	Sales Order #	Salesperson #	Region	Sales	
7	87082	1301	South	$119,423	
8	87083	1302	East	$321,989	
9	87084	1304	East	$155,882	
10	87085	1302	East	$282,097	
11	87086	1302	North	$305,319	
12	87087	1301	East	$111,512	
13	87088	1301	East	$308,051	
14	87089	1304	North	$ 96,069	

Introductio | Sales Report | alesperson Information

Notice that the last "Sales Order #" is 87175, and there are 94 rows of data (row 100 – row 6 = 94 rows).

6	Sales Order #	Salesperson #	Region	Sales
97	87172	1302	West	$154,798
98	87173	1304	North	$353,819
99	87174	1304	East	$256,322
100	87175	1303	North	$ 78,635

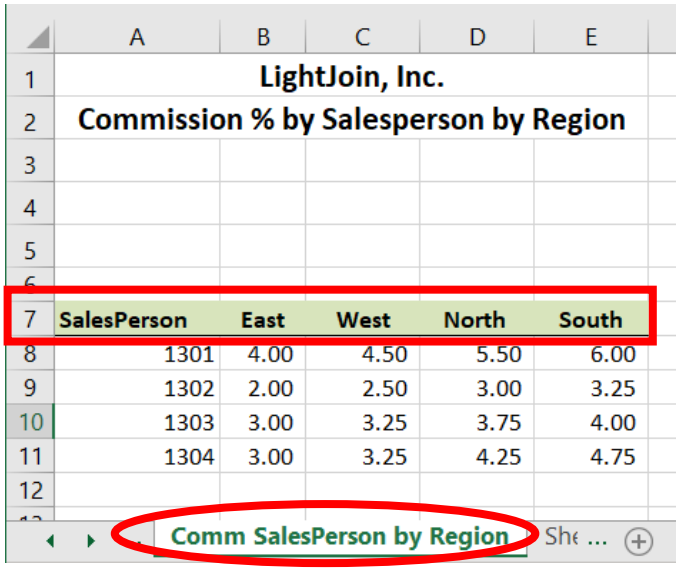
The second stream containing data regarding the Salesperson's name and title is stored in a sheet titled "**Salesperson Information**". Notice that the column headings are in the fourth row.



	A	B	C
1	LightJoin, Inc.		
2	Salesperson Information		
3			
4	SPID	Salesperson	Title
5	1301	George Washington	Sales Associate I
6	1302	John Adams	Senior Sales Associate
7	1303	Thomas Jeferson	Senior Sales Associate
8	1304	James Madison	Sales Associate I
9			

Sales Repor | Salesperson Information | C

The third stream contains the commission percentage earned by the salesperson by region in which the sales were made. This data is stored in a sheet titled "**Comm SalesPerson by Region**". Notice that the column headings are in the seventh row.



	A	B	C	D	E
1	LightJoin, Inc.				
2	Commission % by Salesperson by Region				
3					
4					
5					
6					
7	SalesPerson	East	West	North	South
8	1301	4.00	4.50	5.50	6.00
9	1302	2.00	2.50	3.00	3.25
10	1303	3.00	3.25	3.75	4.00
11	1304	3.00	3.25	4.25	4.75
12					

Comm SalesPerson by Region | She ...

The deliverable consists of two reports. One is a Commission in Dollars Report by Salesperson and the other a Commission in Dollars Report by Salesperson and Region. Both are to be loaded back into Excel.

	A	B
1	Salesperson	Commission Expense
2	George Washington	226,149.84
3	James Madison	274,189.61
4	Thomas Jeferson	127,193.52
5	John Adams	90,935.34
6		
7	Total	718,468.31

Mark Friedman
University of Miami

	A	B	C	D	E	F
1						
2						
3	**Sum of Commission Expense**	**Column Labels**				
4	**Row Labels**	**East**	**North**	**South**	**West**	**Grand Total**
5	George Washington	84,829.28	29,372.96	82,074.30	29,873.30	226,149.84
6	James Madison	60,005.82	136,168.64	44,940.09	33,075.06	274,189.61
7	John Adams	17,192.18	26,615.28	26,052.39	21,075.49	90,935.34
8	Thomas Jeferson	30,874.29	52,700.71	15,502.96	28,115.56	127,193.52
9	**Grand Total**	**192,901.57**	**244,857.59**	**168,569.74**	**112,139.41**	**718,468.31**

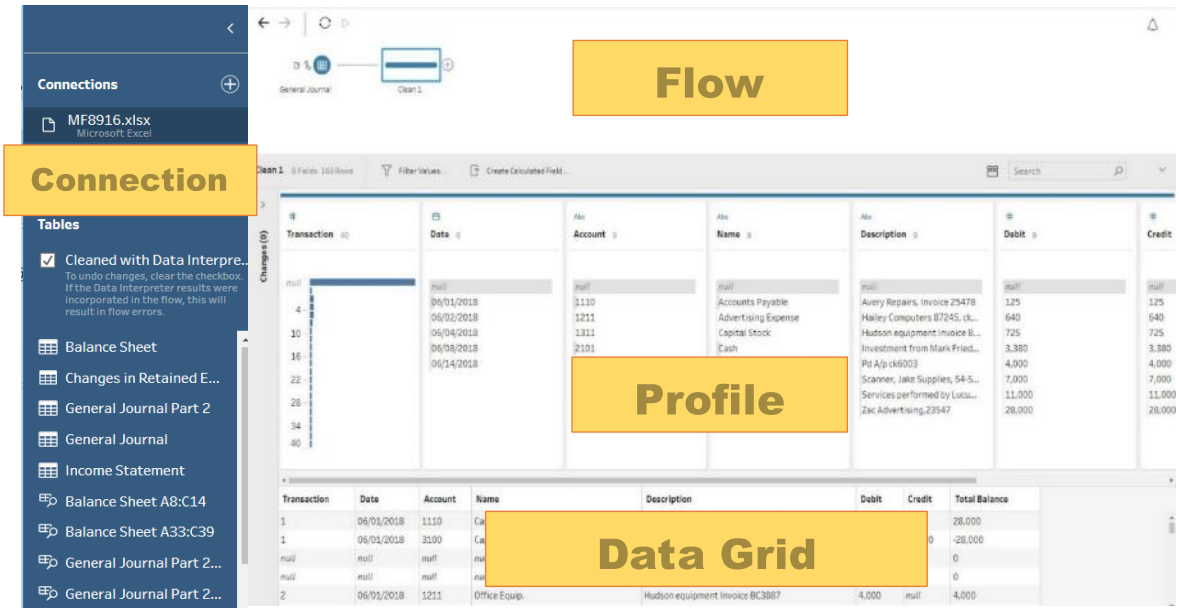
Mark Friedman
University of Miami

Start Tableau Prep

Note: You cannot extract data from an open Excel file, so make sure that the Excel file is closed before starting to input the data.

The Tableau Prep Builder workspace consists of the Connections pane and three coordinated areas that help you interact with and explore your data:

- Flow pane: A visual representation of your operation steps as you prepare your data.
- Profile pane: A summary of each field in your data sample. See the shape of your data and quickly find outliers and nulls.
- Data grid pane: The row level detail for your data.

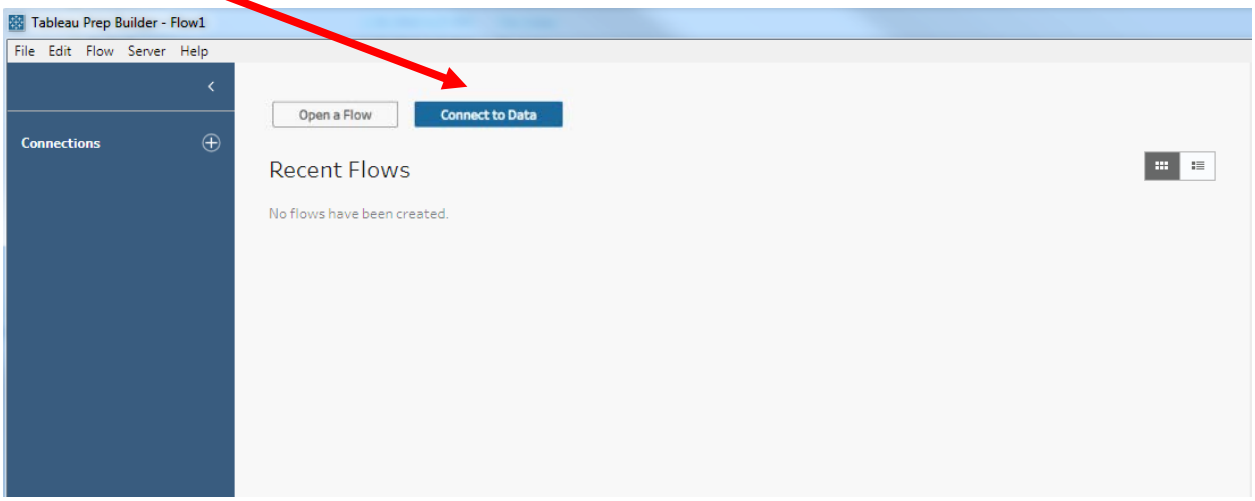


After you connect to your data and begin building your flow, you add steps in the Flow pane. These steps function as a lens into the structure of your data, as well as a summary of operations that is applied to your data. Each step represents a different category of operations that you define.

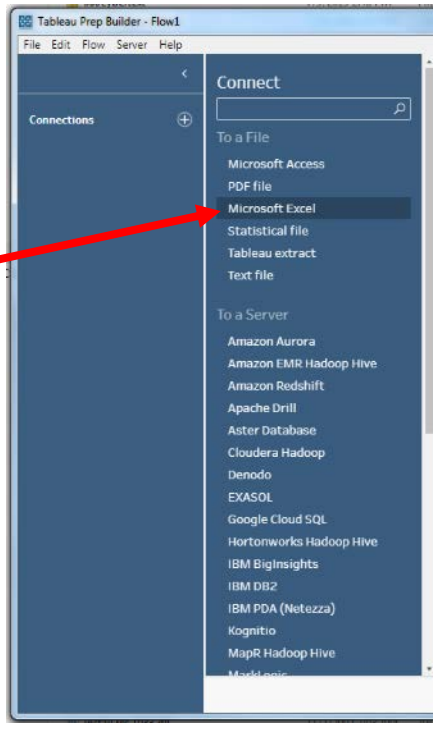
In the Configuration area, select the “Connect a File or Database” drop-down arrow.

Step 1: Input Excel file

To open Excel file via Tableau Prep, press green button “Connect to Data” on the Flow pane.

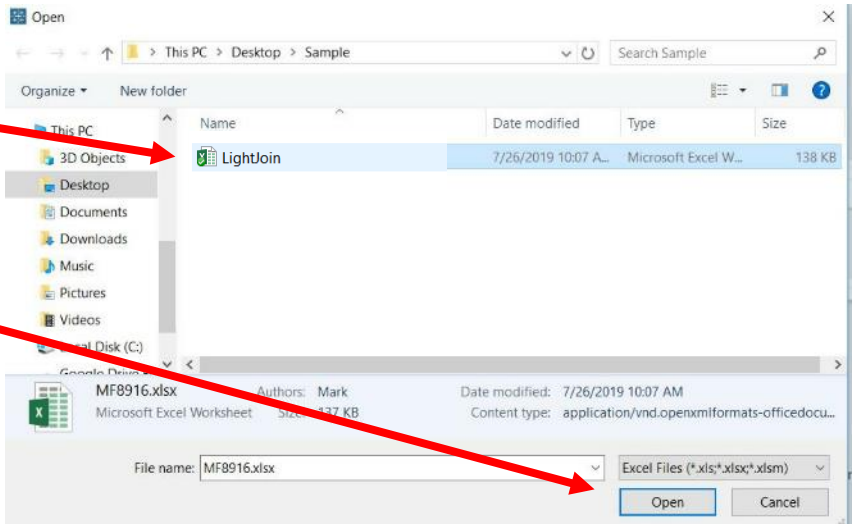


On the connections pane select Microsoft Excel.



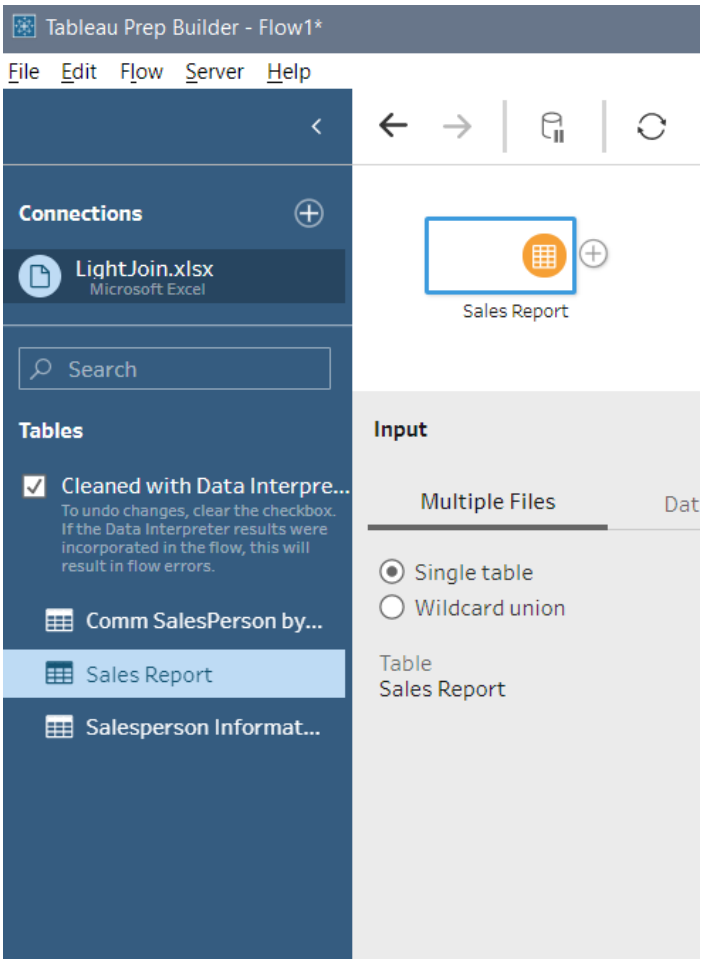
Select the file.

Select “Open”.

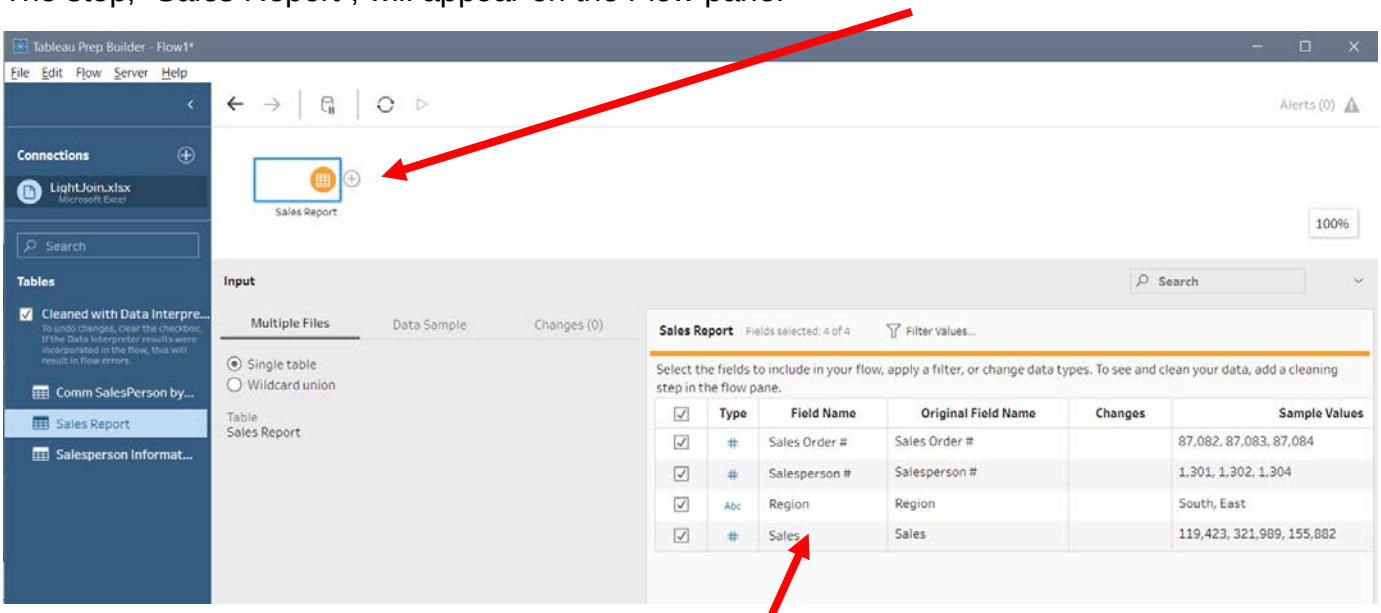


Remember that the field names are in the sixth row. Check “Cleaned with Data Interpreter” tick box to let Tableau Prep adjust for the input data.

Select the “Sales Report” sheet on the Connections pane by either double click on the sheet name or drag the sheet name to the canvas.



The step, “Sales Report”, will appear on the Flow pane.

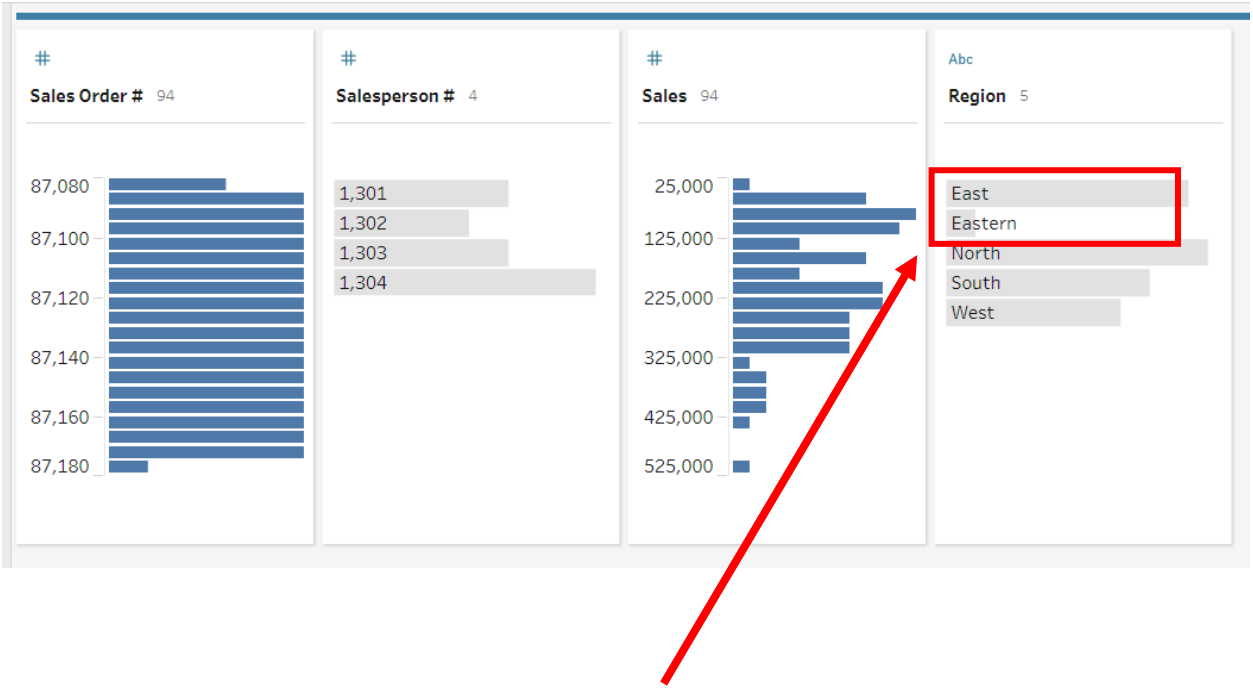


The sample data will be visible on the Profile pane.

Step 2: Clean the data

Review the data that we just added. Click the plus sign and add “Clean Step”.

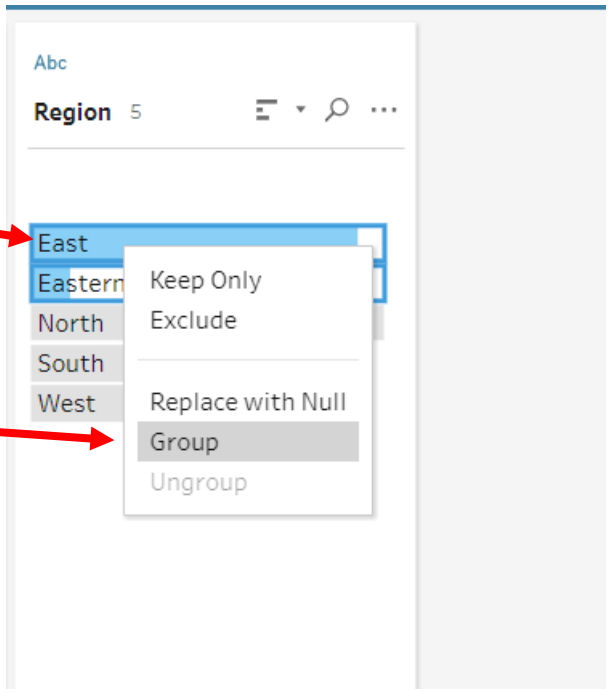
On the profile pane review the data.



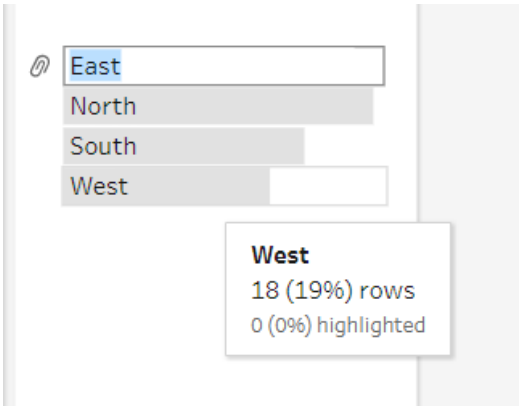
In the “Region” column two regions looks similar, but not the same: “East” and “Eastern”. In one or more records “East” was incorrectly entered as “Eastern”. We need to correct this manual input error.

Group both values and give the group a common name. Select the two regions by first clicking on "East" then hold down the control key click on the second field "Eastern" and then release the control key (Ctr+).

While hovering over the selected items right click and select "Group".

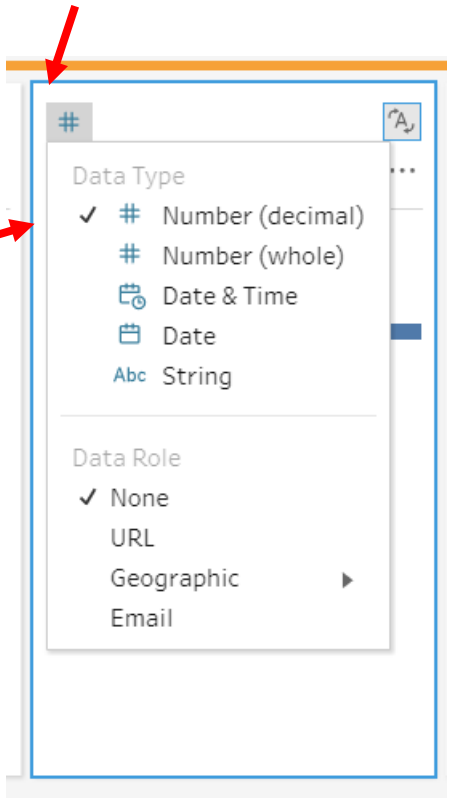


If necessary, name the group "East"



Mark Friedman
University of Miami

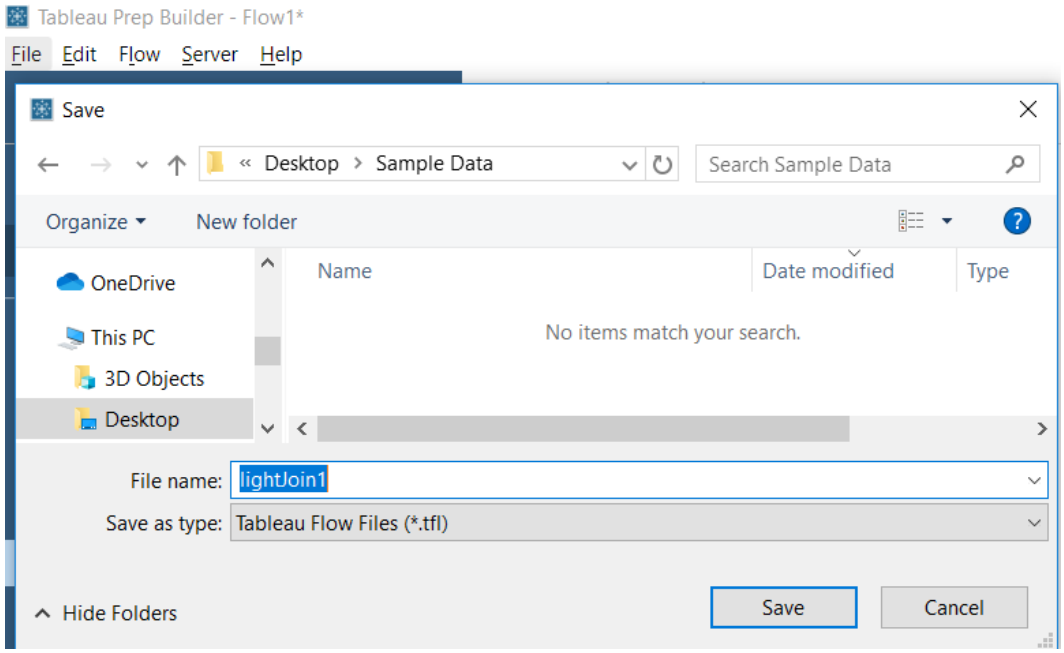
To assure that Tableau calculates math functions correctly change the data type of the "Sales" field to Number (decimal). Select the "#" button on top of the column and select "# Number (decimal)"



Save the workflow regularly.

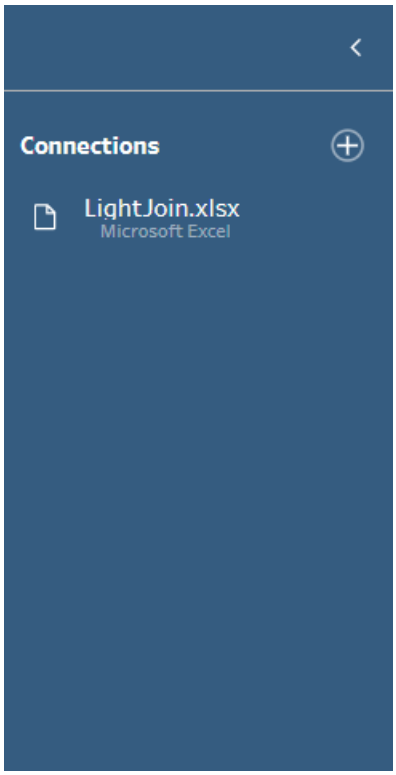
Select “File”

Enter the “File name:” and Save.



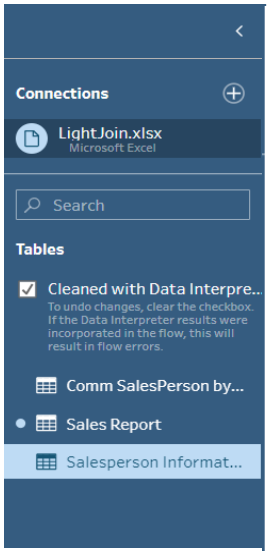
Step 3: Input "Salesperson Information" sheet

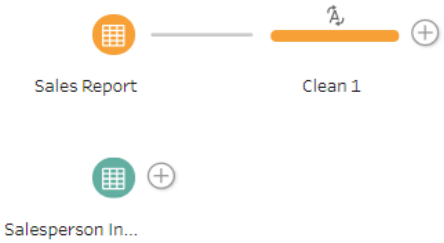
Select "LightJoin.xls" from the "Connections" panel.



Remember that the field names are in the fourth row. The "Cleaned with Data Interpreter" should still be selected.

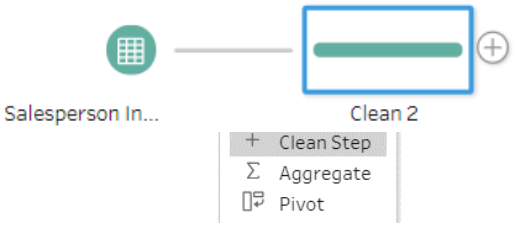
Select the "Salesperson Information" sheet on the Connections pane by either double click on the sheet name or drag the sheet name to the canvas.

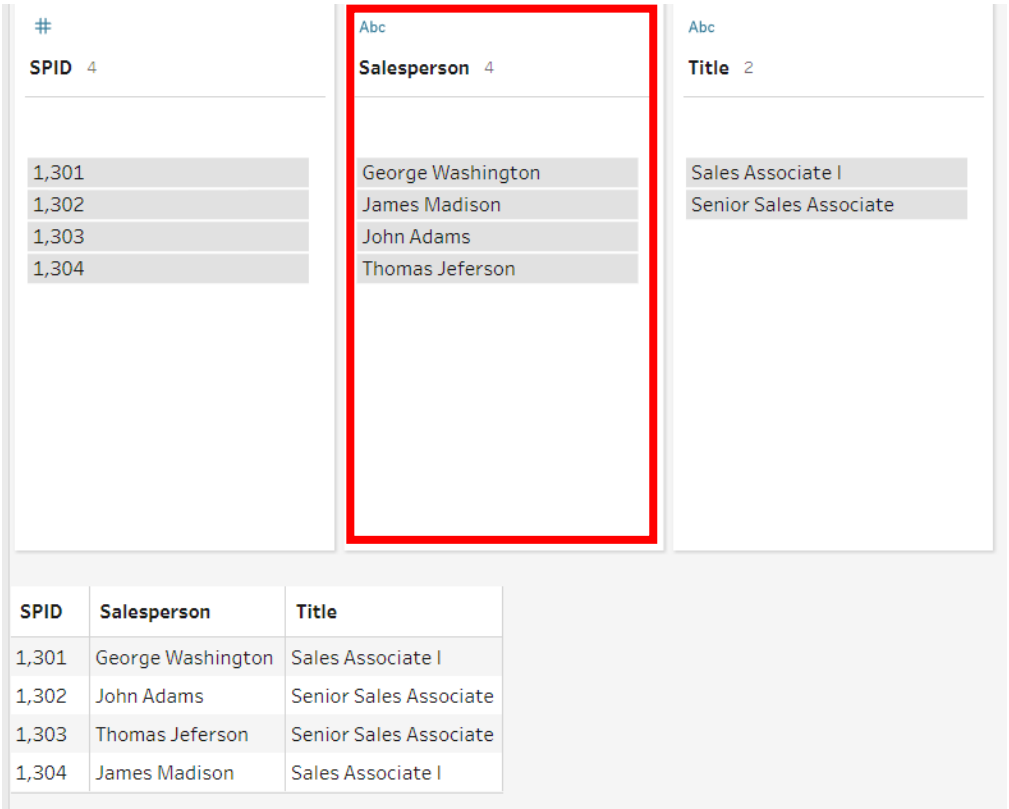




Step 4: Clean the data

Review the data that we just added. Click the plus sign and add step “Clean 2”.

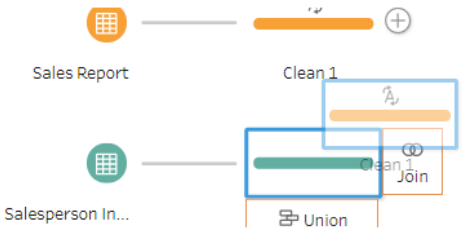




SPID	Salesperson	Title
1,301	George Washington	Sales Associate I
1,302	John Adams	Senior Sales Associate
1,303	Thomas Jeferson	Senior Sales Associate
1,304	James Madison	Sales Associate I

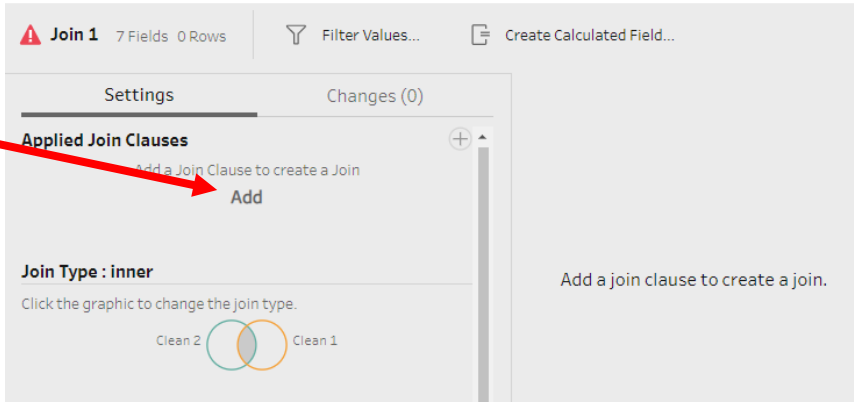
Step 5: Combine Data from first Two Streams on a Common Field

Now join the two the streams. Drag the "Clean 1" step from the first stream to the "Clean 2" step in the second stream and a Join box will appear. Drop the "Clean 1" step into the Join box.

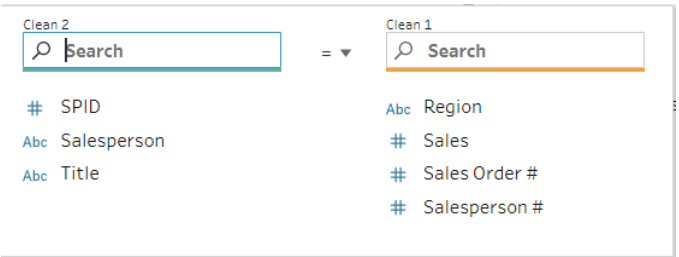


As a result, step "Join 1", a combination of the two two streams will appear. Select the "Join 1" step.

Select "Add" in the "Applied Join Clauses" section.

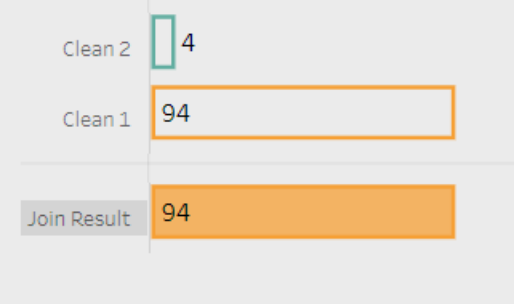


Select the fields to join on, “SPID” from Cases 2 and “Salesperson #” from “Case 1” .



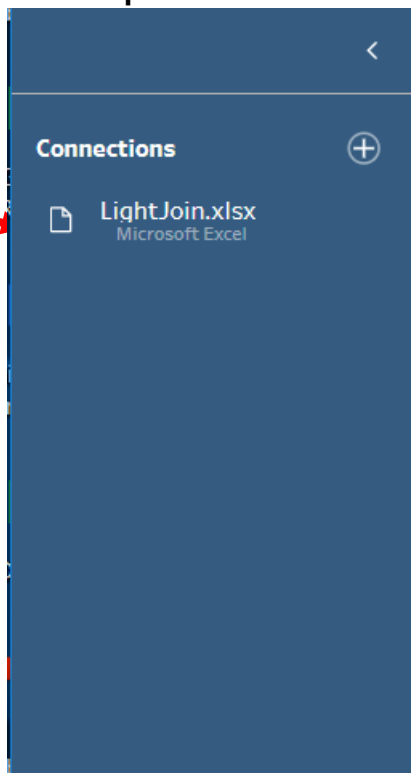
The default “Join Type” is “inner” which does not have to be changed.

The join was successful as 94 rows were combined.



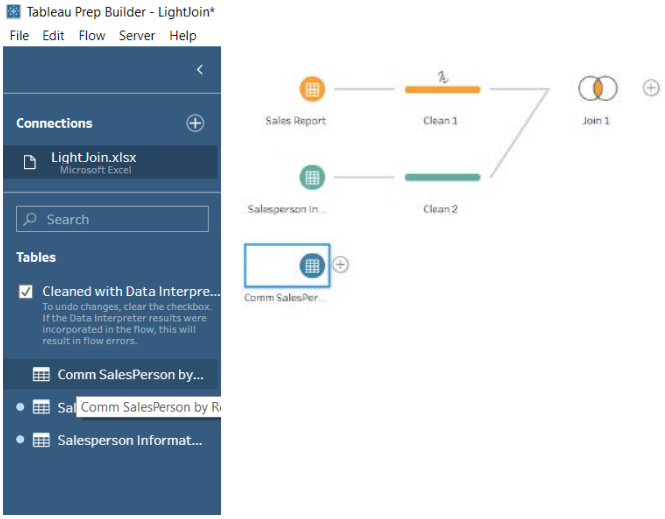
Step 6: Input "Comm SalesPerson by Region" sheet and pivot data

Select "LightJoin.xls" from the "Connections" panel.

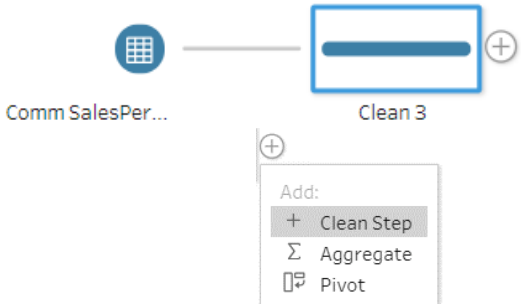


Remember that the field names are in the seventh row. The "Cleaned with Data Interpreter" should still be selected.

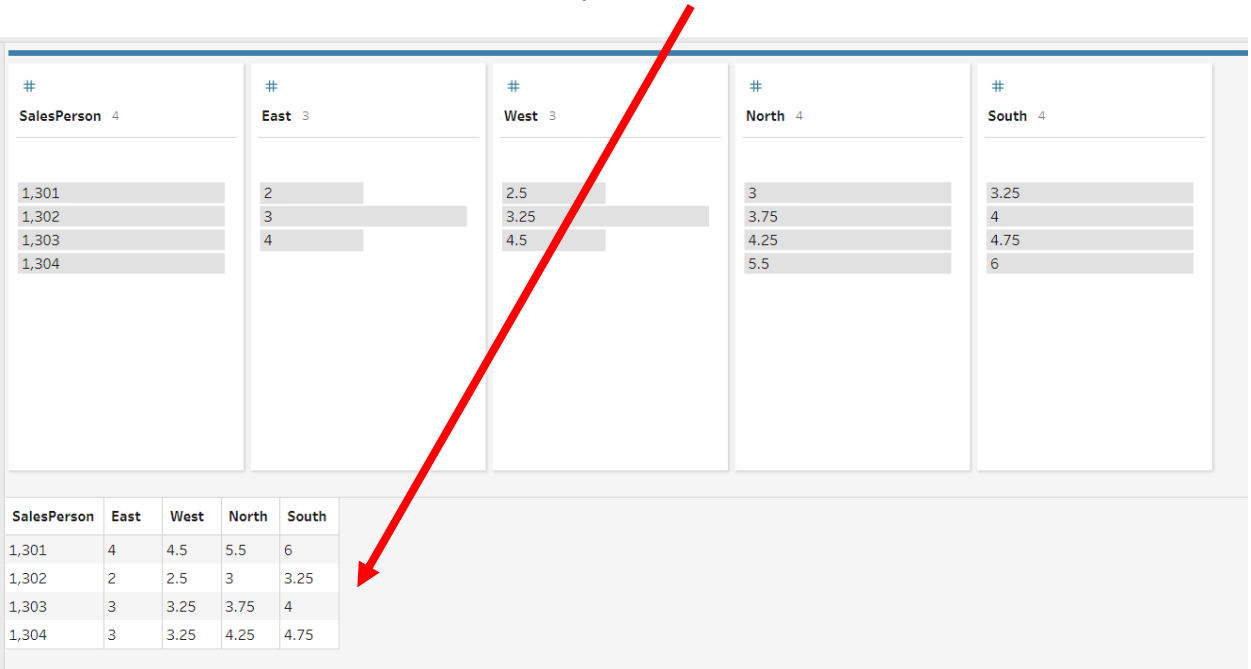
Select the "**Comm SalesPerson by Region**" sheet on the Connections pane by either double click on the sheet name or drag the sheet name to the canvas.



Step 7: Clean the data
Review the data that we just added. To do that click plus sign and add step "Clean 3".

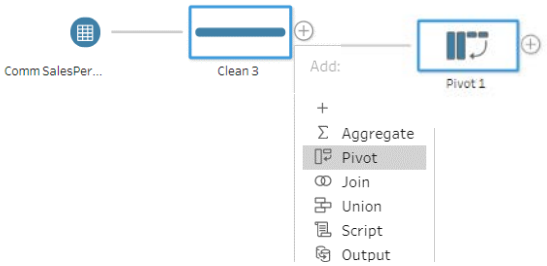


Data in a matrix format cannot be joined to the previous inputs in the current format. Each record must be converted into multiple records.

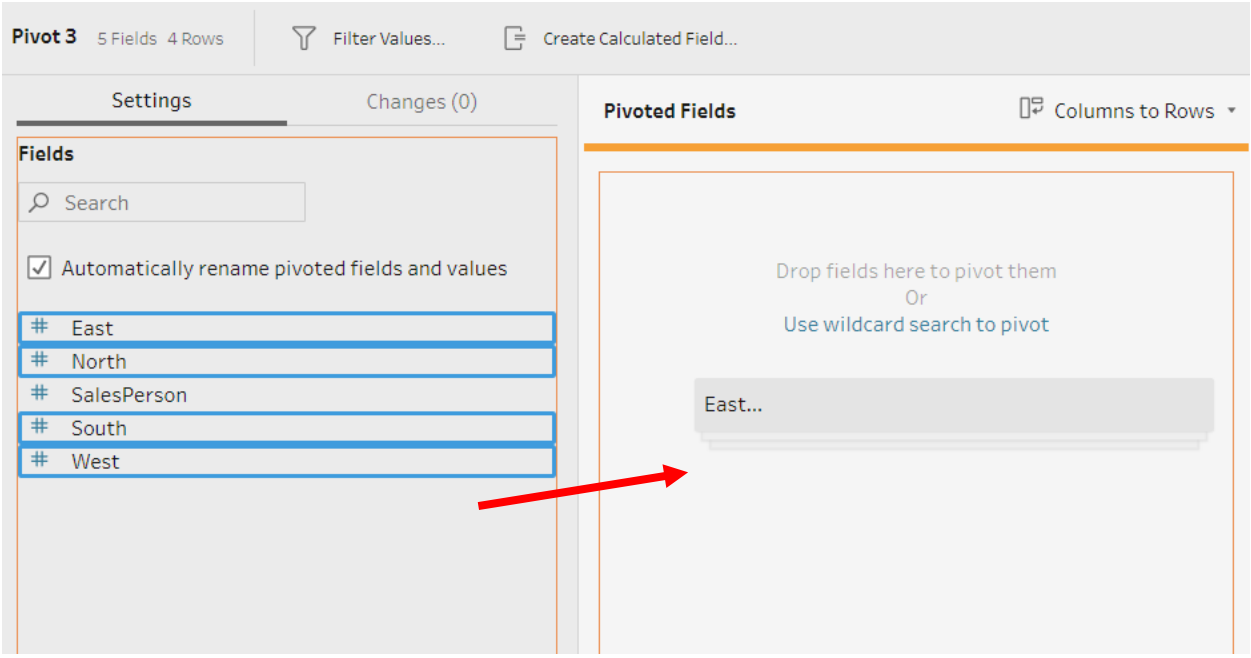


Step 8: Pivot the data.

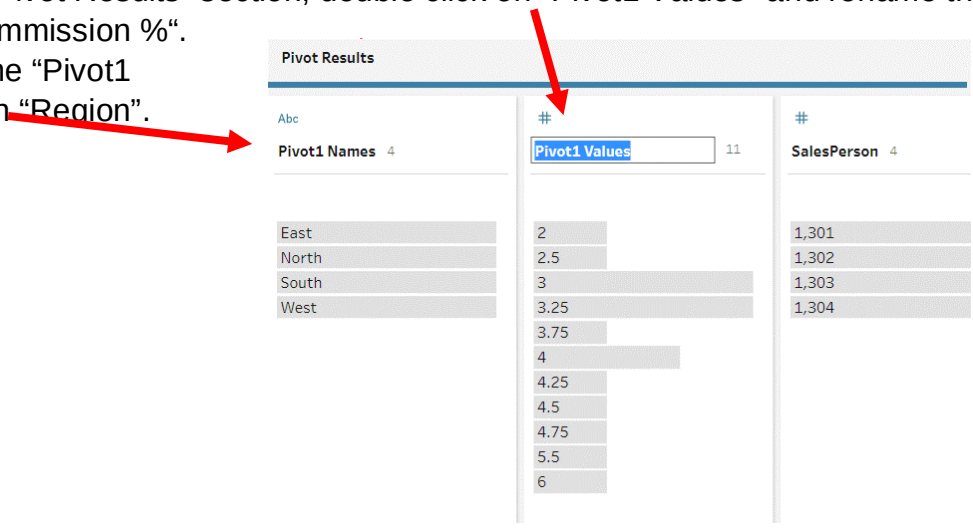
Add a “Pivot” step to the “Comm SalesPerson by Region” steam’s “Clean 3” step.

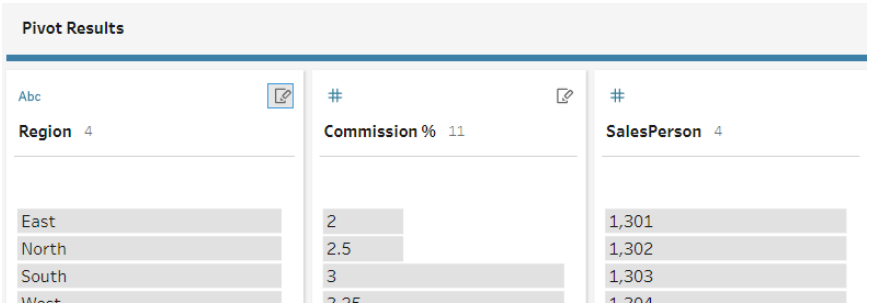


Select all four regions columns and drag them to Pivoted Fields Area



Select the “Pivot Results” section, double click on “Pivot1 Values” and rename the column “Commission %“.
Then rename “Pivot1 Names” with “Region”.

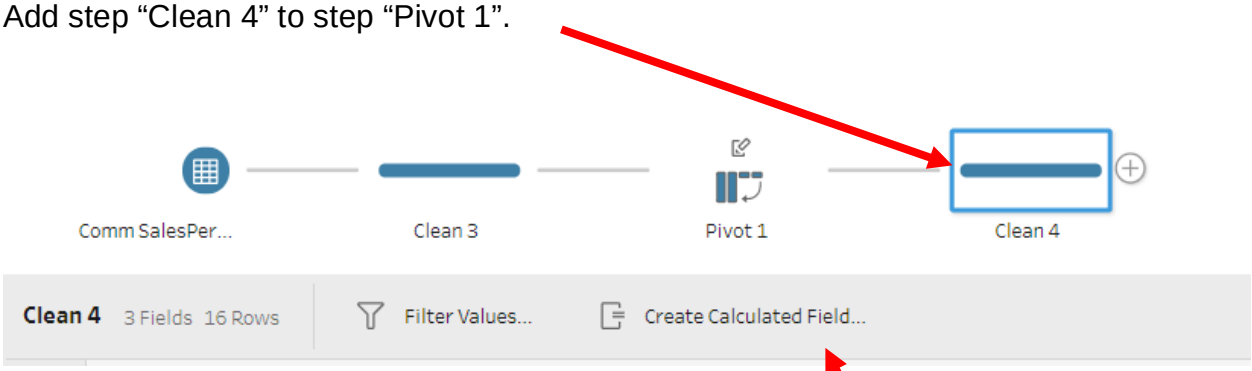




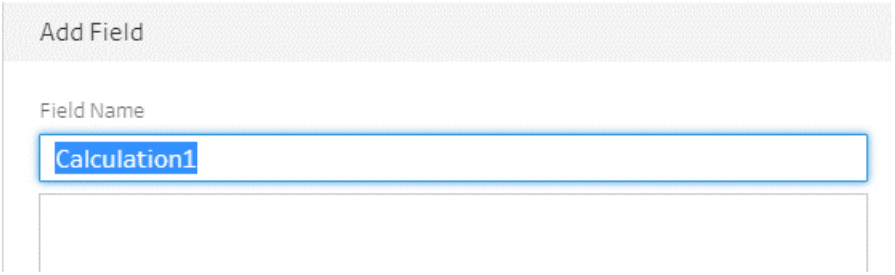
To combine two streams of data there must be a unique key. That unique field is referred to as the primary key in the one side of the join. The primary field will be created by combining the “Salesperson” field with a “.” and the “Region” field. The primary key for the first record will be “1301.East”. Since a numeric field “Salesperson” is being combined with a string field, “Region” the numeric field must be converted to a string field.

Step 9: Create a Primary Key Field

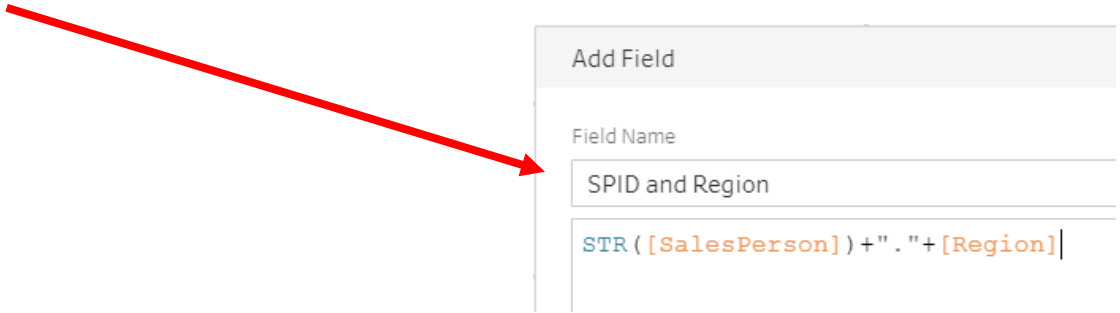
Add step “Clean 4” to step “Pivot 1”.



To create a Primary Key field, select “Create Calculated Field”.

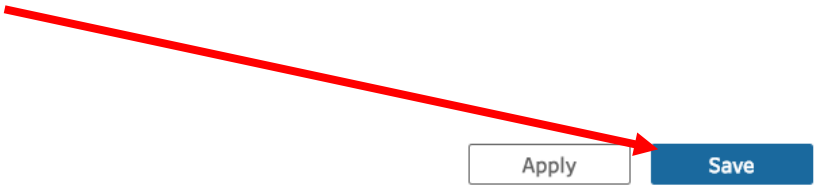


Change the “Field Name” to “SPID and Region”.

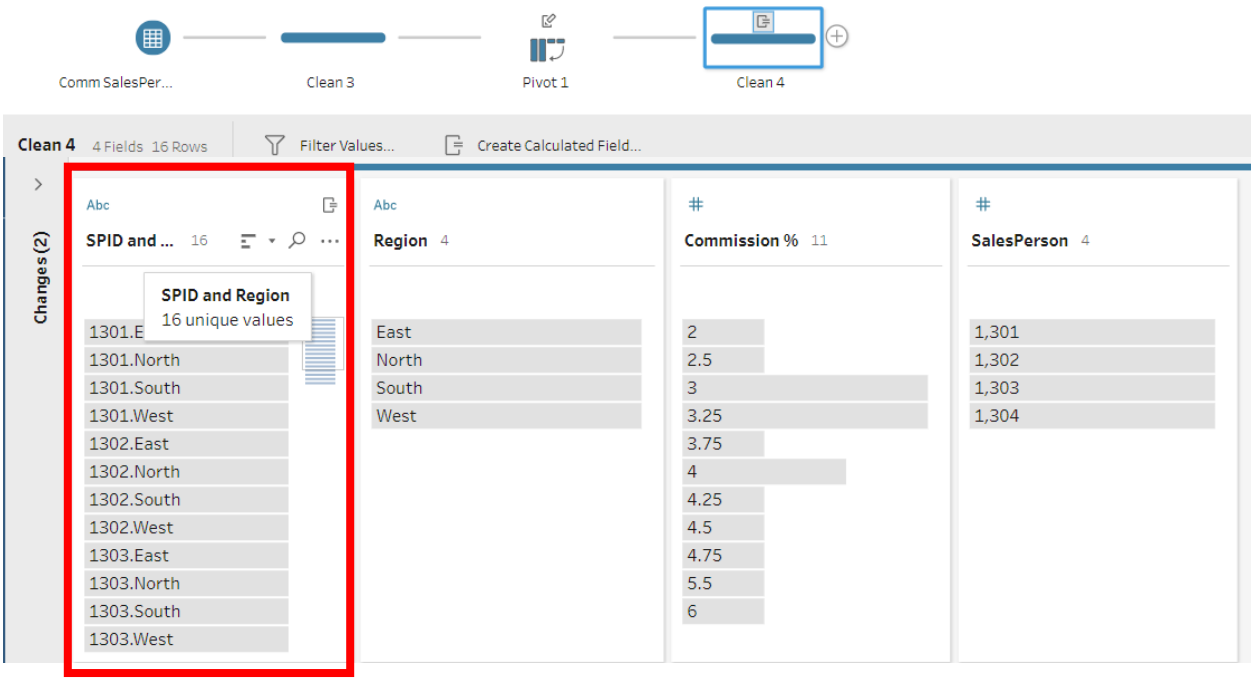


Type: **STR([SalesPerson]) + "." + [Region]**

then click Save.



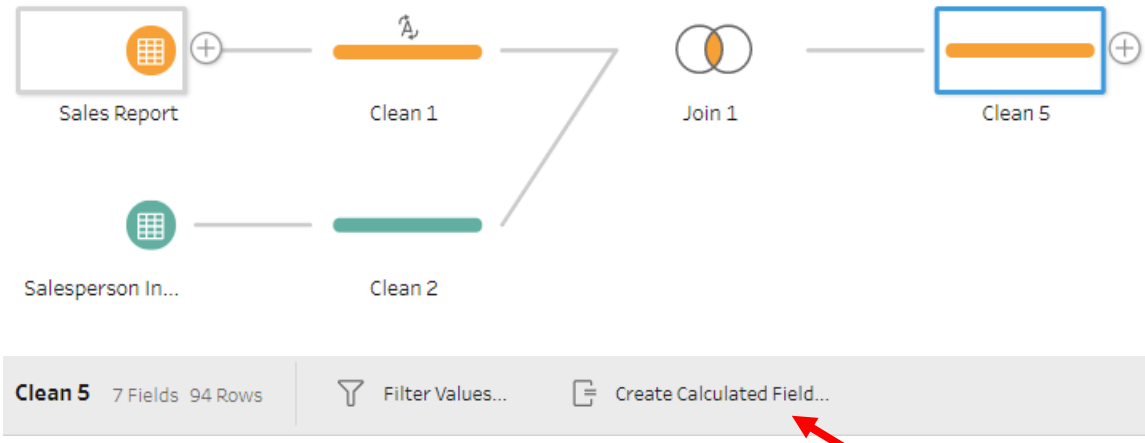
The “SPID and Region” field has been populated.



Step 10: Create a Foreign Key Field

The combined first and second stream of data also needs a field that combines the "SPID" and the "Region" in the same exact format that was used in the third stream. This new field is referred to as a foreign key and will be used to link to the data in the third stream.

Add step "Clean 5" to the joined first and second stream's "Join 1" step.



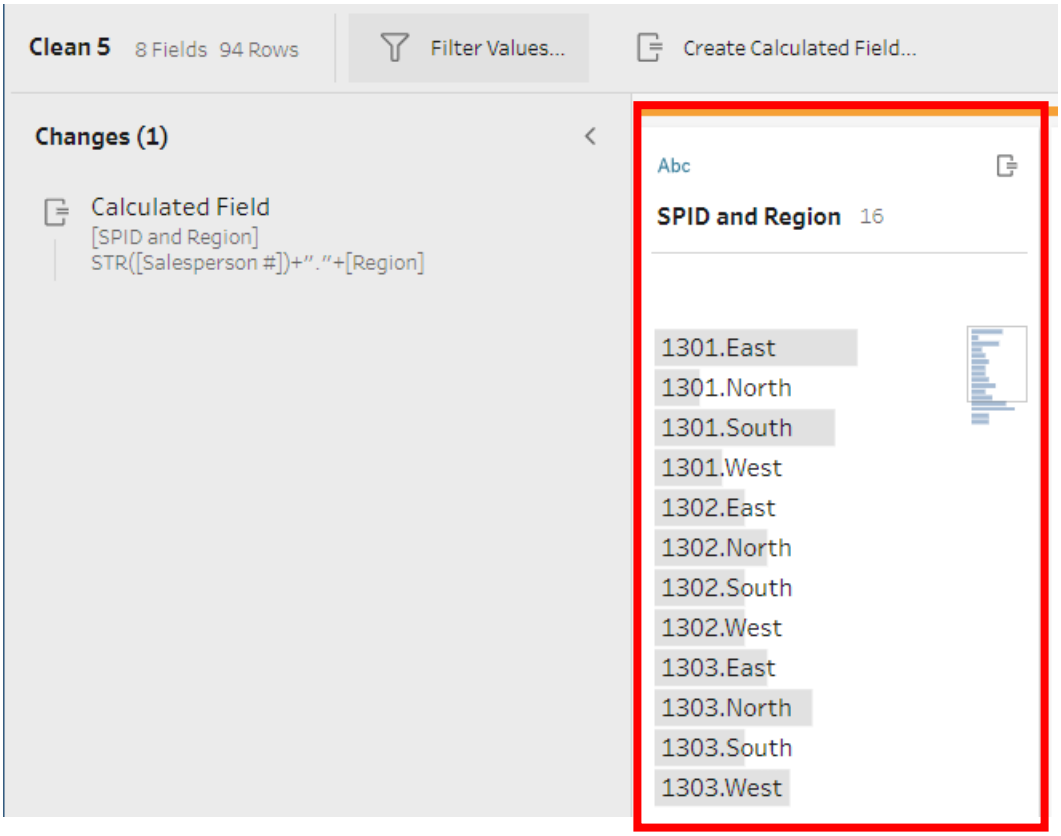
To create a Foreign Key field, select "Create Calculated Field".

Change the "Field Name" to "SPID and Region".

Type: **STR([SalesPerson #]) + "." + [Region]**

then click Save.

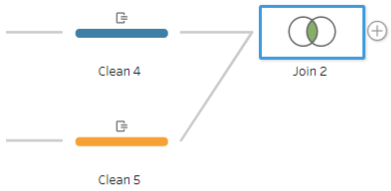
The “SPID and Region” field has been populated with 16 different values for “SPID and Region”.



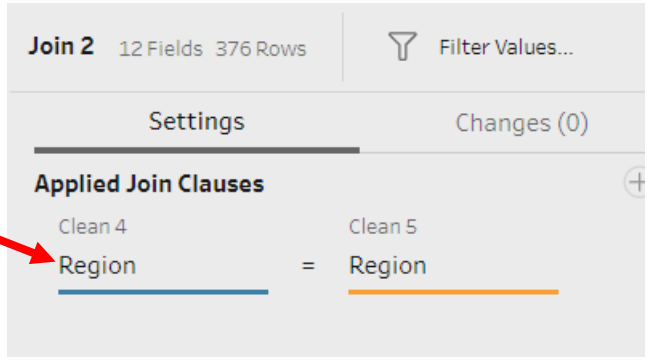
Step 11: Combine First and Second Stream of Data to the Third Stream of Data

Now that we have a Primary key and a Foreign key, we can join the third stream of data with the combined first two streams. Drag the “Clean 5” step from the combined first and second stream to the “Clean 4” step in the third stream. Drop the “Clean 5” step into the Join box.

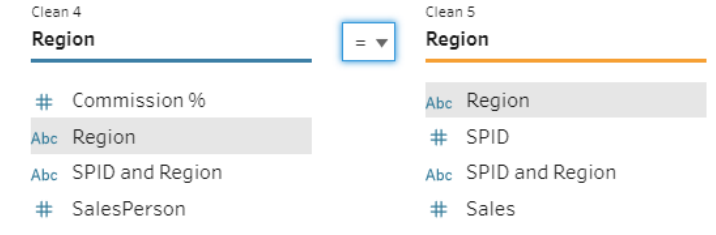
Step “Join 2” will appear. Select the “Join 2” step.

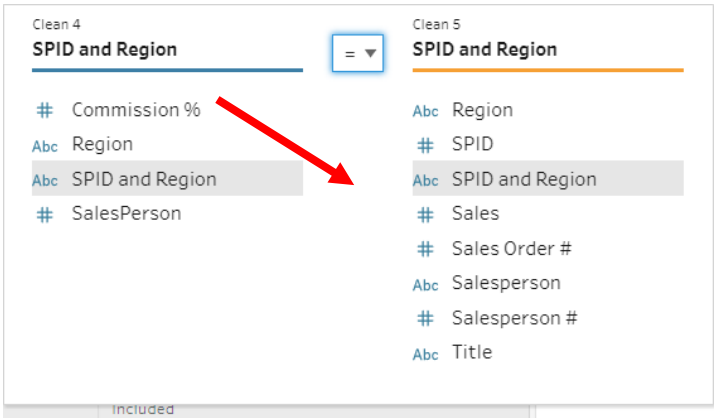


Select “Region in the “Applied Join Clauses” section.



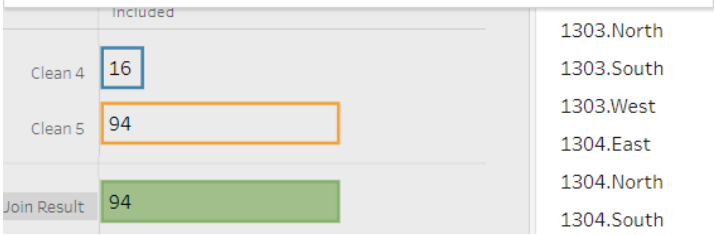
Change the fields from Region to “SPID and Region” in both “Clean 4” and “Clean 5” steps.

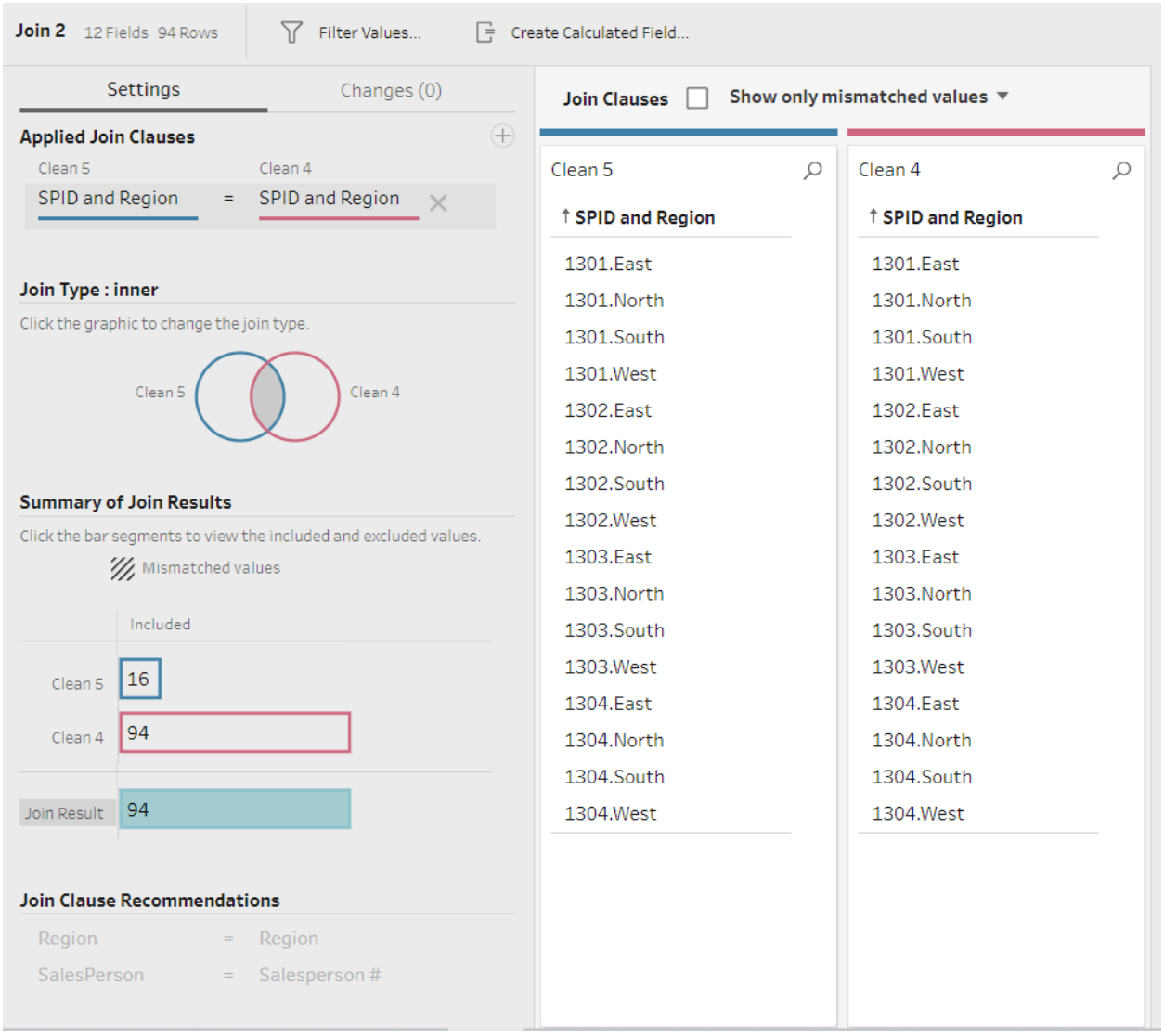




The default “Join Type” is “inner” which does not have to be changed.

The join was successful as 94 rows were combined.

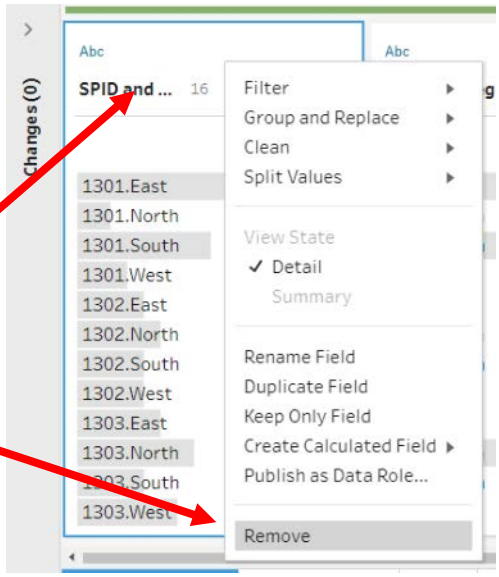



Join 2 12 Fields 94 Rows
Filter Values...
Create Calculated Field...
Settings
Changes (0)
Applied Join Clauses
Clean 5
Clean 4
SPID and Region = SPID and Region
Join Type : inner
Click the graphic to change the join type.
Clean 5
Clean 4
Summary of Join Results
Click the bar segments to view the included and excluded values.
Mismatched values
Included
Clean 5 16
Clean 4 94
Join Result 94
Join Clause Recommendations
Region = Region
SalesPerson = Salesperson #
Join Clauses
Show only mismatched values
Clean 5
SPID and Region
1301.East
1301.North
1301.South
1301.West
1302.East
1302.North
1302.South
1302.West
1303.East
1303.North
1303.South
1303.West
1304.East
1304.North
1304.South
1304.West
Clean 4
SPID and Region
1301.East
1301.North
1301.South
1301.West
1302.East
1302.North
1302.South
1302.West
1303.East
1303.North
1303.South
1303.West
1304.East
1304.North
1304.South
1304.West

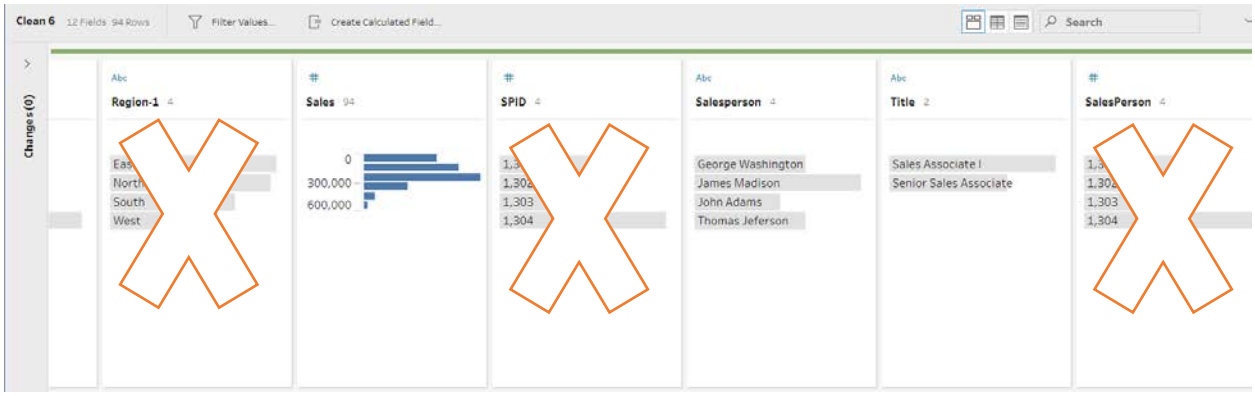
Step 12: Remove Fields and Calculate the Commission on Every Sales Order

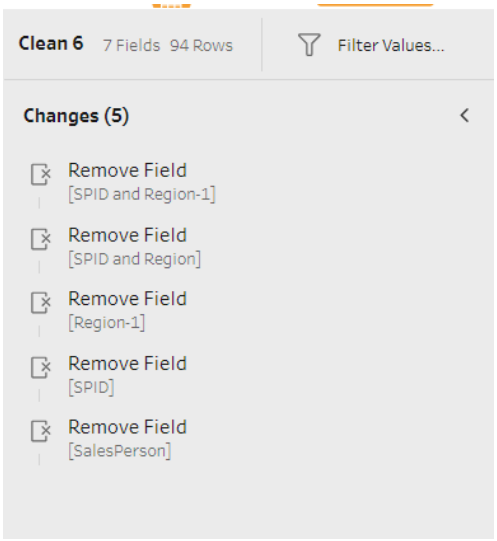
Add a step “Clean 6” after the step “Join 2” and remove some duplicating fields.

To remove a field right click the column header and select Remove.



- SPID and Region-1
- SPID and Region
- Region-1
- SPID
- SalesPerson (The number not the name)





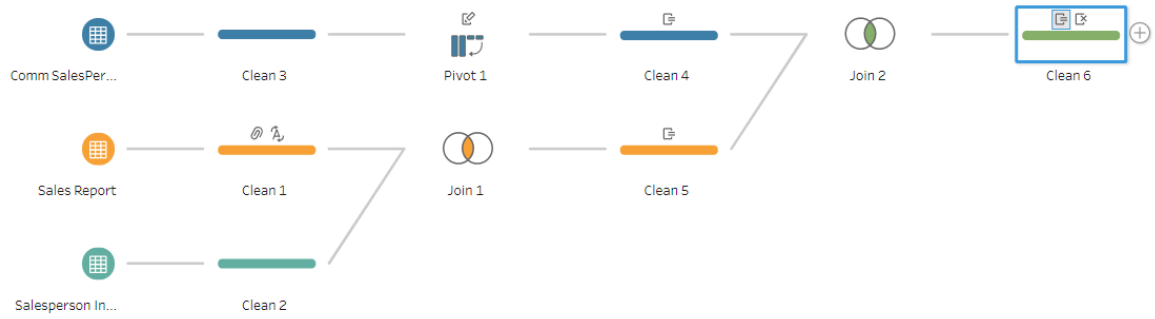
Now calculate the commission on every sales order.

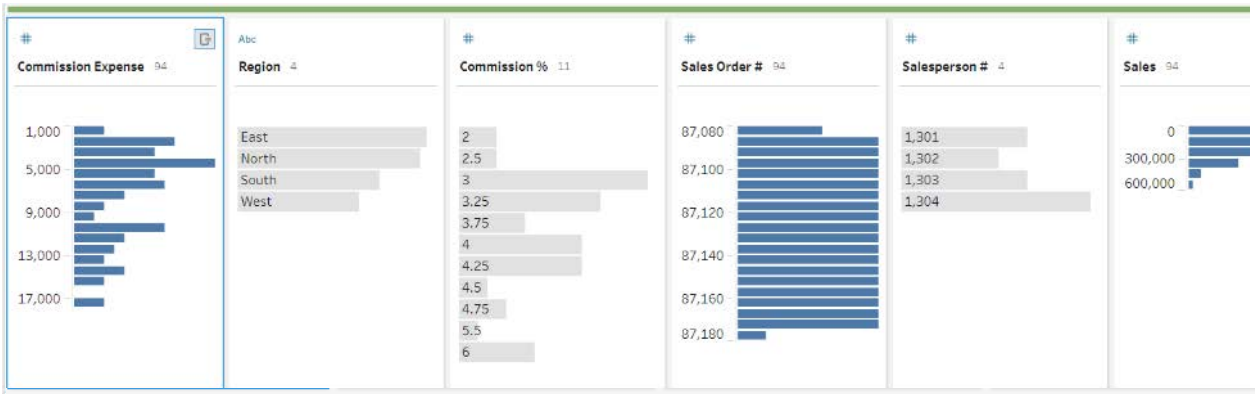
Click on “Create Calculated Field”

Enter “Commission Expense” as the “Field Name”.

Type the formula Round(**[Sales]*[Commission %]/100,2)**

then click Save.

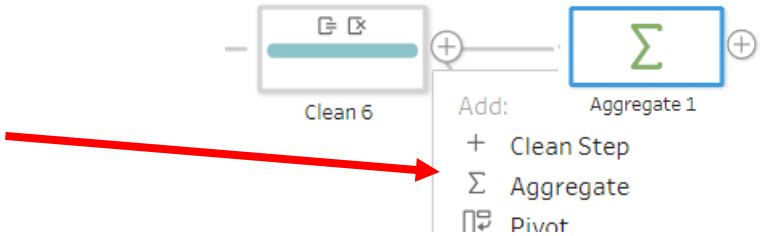




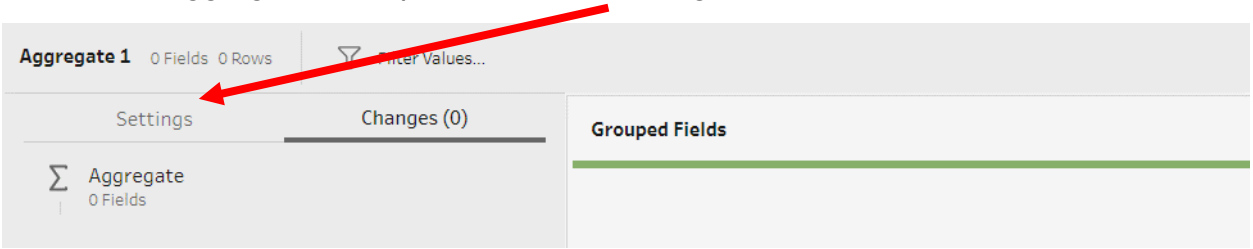
Commission Expense	Region	Commission %	Sales Order #	Salesperson #	Sales	Salesperson	Title
7,165.38	South	6	87,082	1,301	119,423	George Washing	Sales Associate I
6,439.78	East	2	87,083	1,302	321,989	John Adams	Senior Sales Associa
4,676.46	East	3	87,084	1,304	155,882	James Madison	Sales Associate I
5,641.94	East	2	87,085	1,302	282,097	John Adams	Senior Sales Associa
9,159.57	North	3	87,086	1,302	305,319	John Adams	Senior Sales Associa
4,460.48	East	4	87,087	1,301	111,512	George Washing	Sales Associate I
12,322.04	East	4	87,088	1,301	308,051	George Washing	Sales Associate I

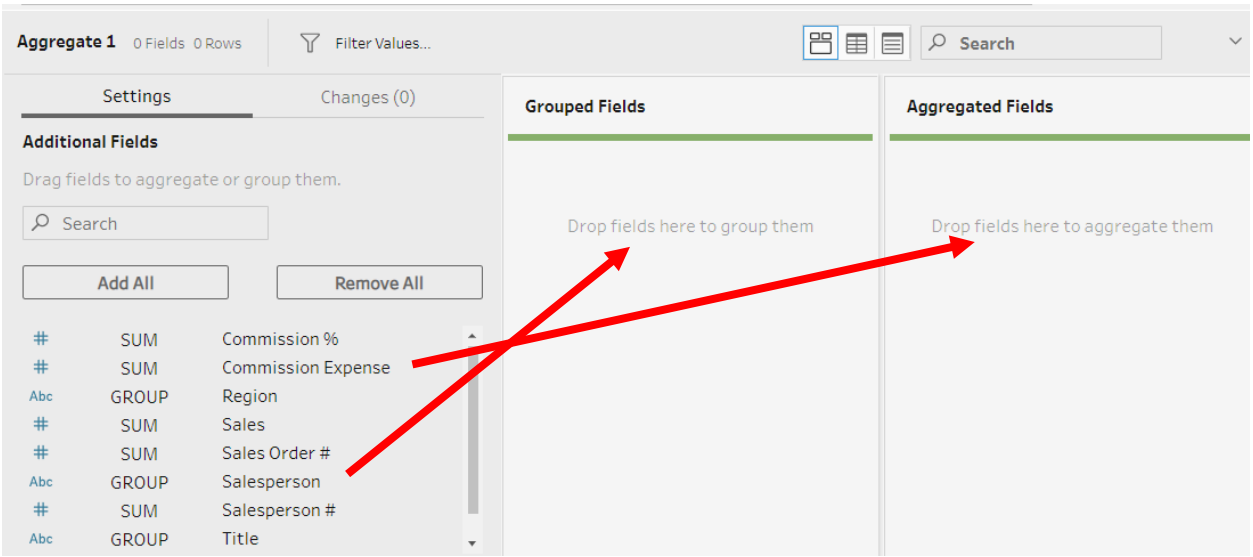
Step 13: Create the "Commission Expense by Salesperson" Report

Add a step "Aggregate 1" after the step "Clean 6 ".



Select the "Aggregate 1" step, then select "Settings".





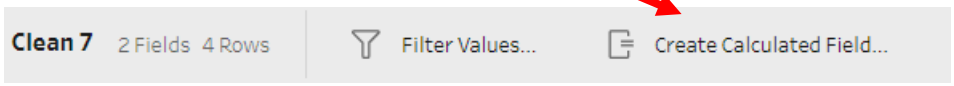
Drag the "Salesperson" field to the "Grouped Fields" area and "Commission Expense" field to the "Aggregated Fields" area.

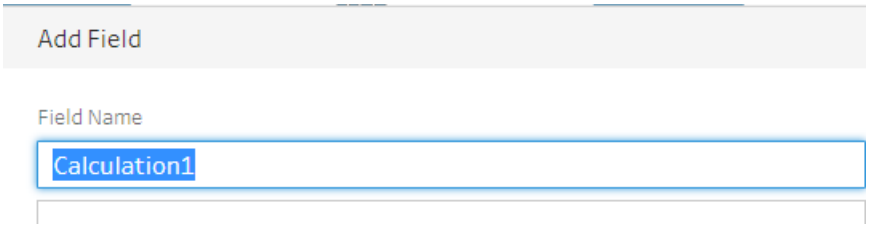
The decimal places in the “Commission Expense” fields are not uniform.

Step 14: Round the Commission Expense

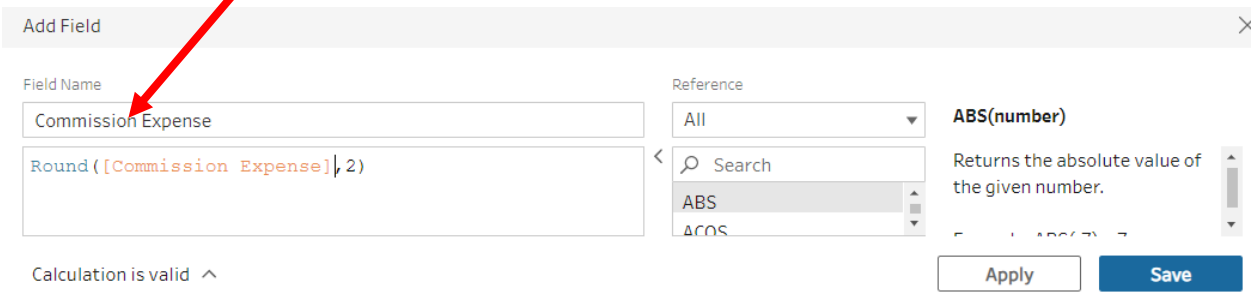
Add a step “Clean 7” after the step “Aggregate 1”.

Select “Create a Calculated Field”



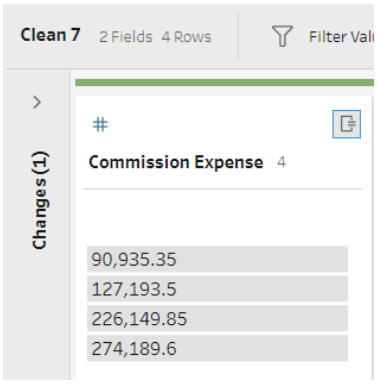


Put the result back in the “Commission Expense” field by naming the field “Commission Expense”.



Type in the formula, Round([Commission],2) and select “Save”.

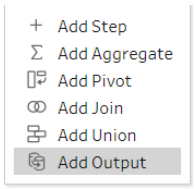
The data in the commission expense field now has two decimal places.



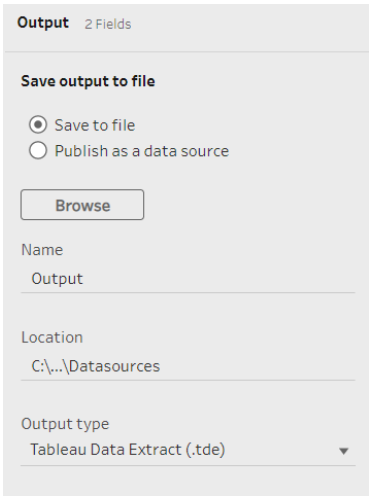
Step 15: Export Report to Excel

Add a step “Output” after the step “Clean 7”.

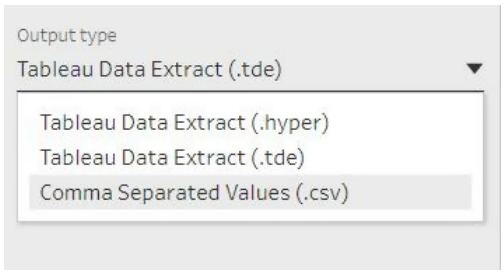
Select the step “Output”



Select “Output type”

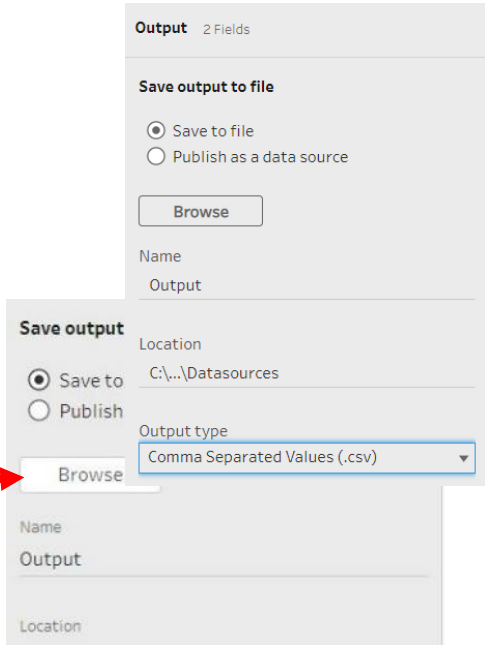


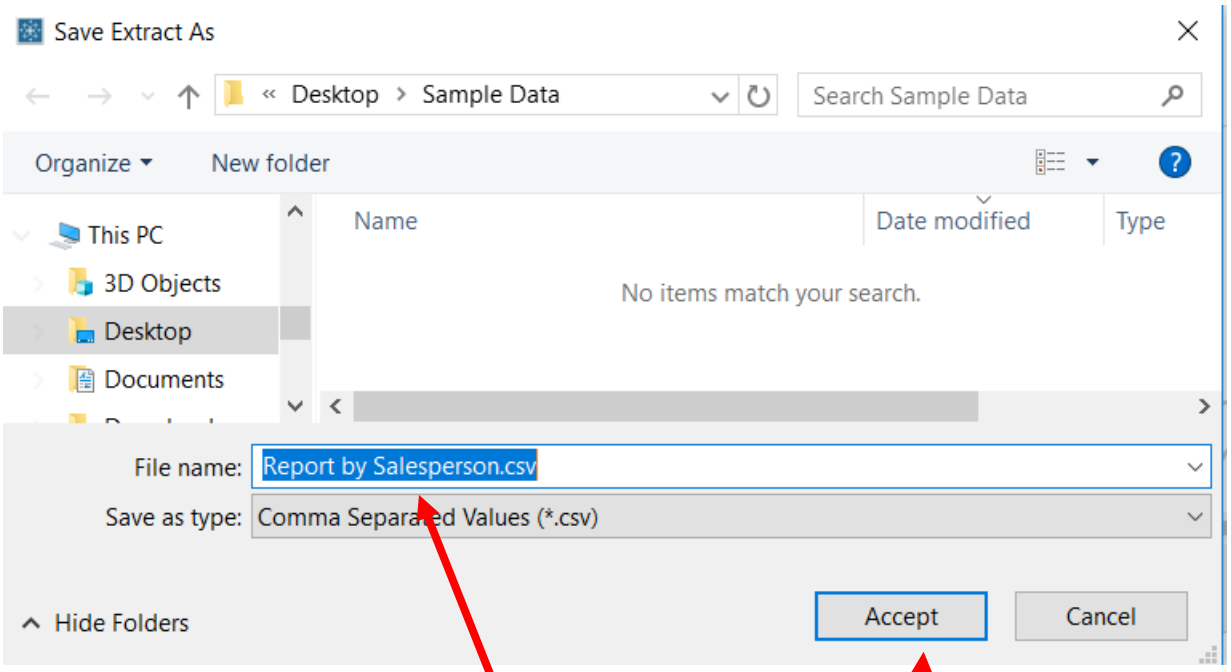
Select the .csv file type.



Select “Save to file” option button.

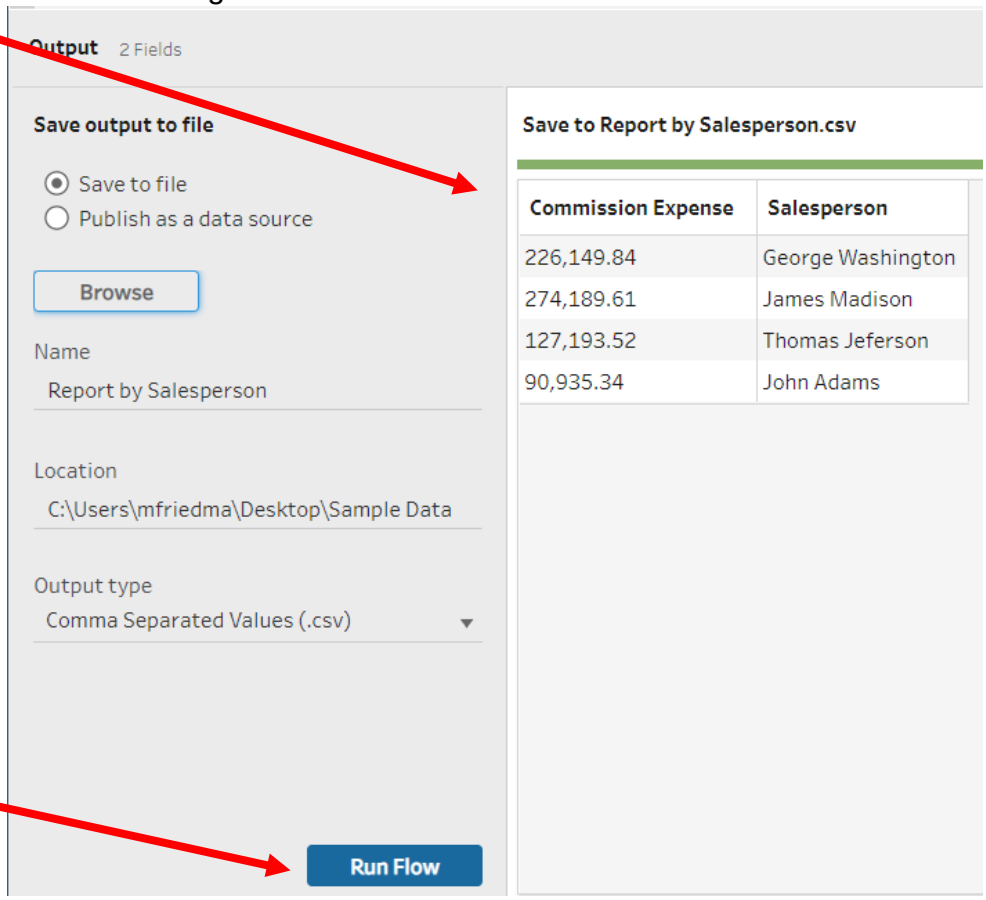
Select “Browse” button





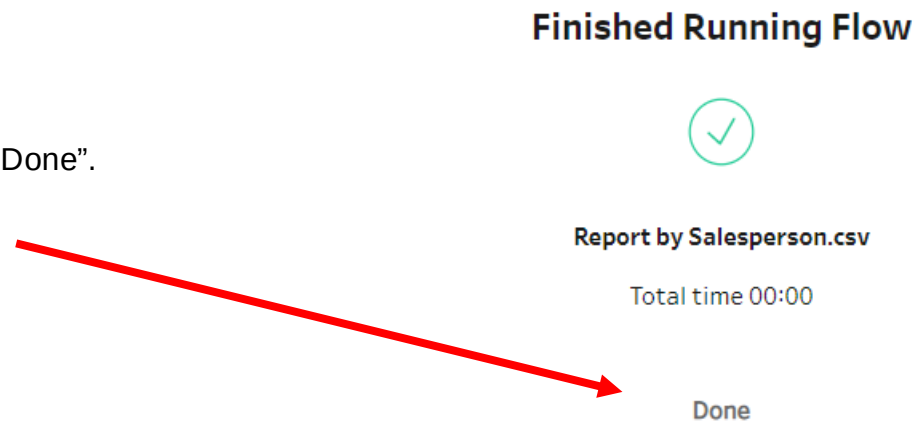
Select the location for the file, specify the file name and select “Accept”.

The preview shows the data that will be generated.

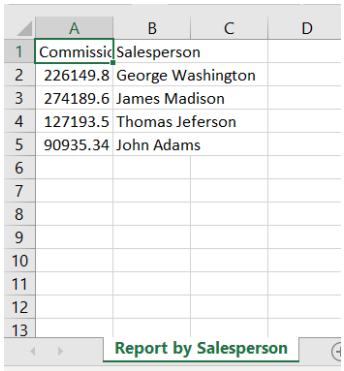


Select “Run Flow”

Select “Done”.

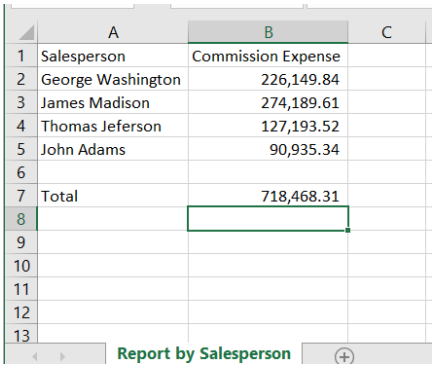


Navigate to the location and double click to open the CSV file you just saved. Make the change that are required in Excel.



	A	B	C	D
1	Commissic	Salesperson		
2	226149.8	George Washington		
3	274189.6	James Madison		
4	127193.5	Thomas Jeferson		
5	90935.34	John Adams		
6				
7				
8				
9				
10				
11				
12				
13				

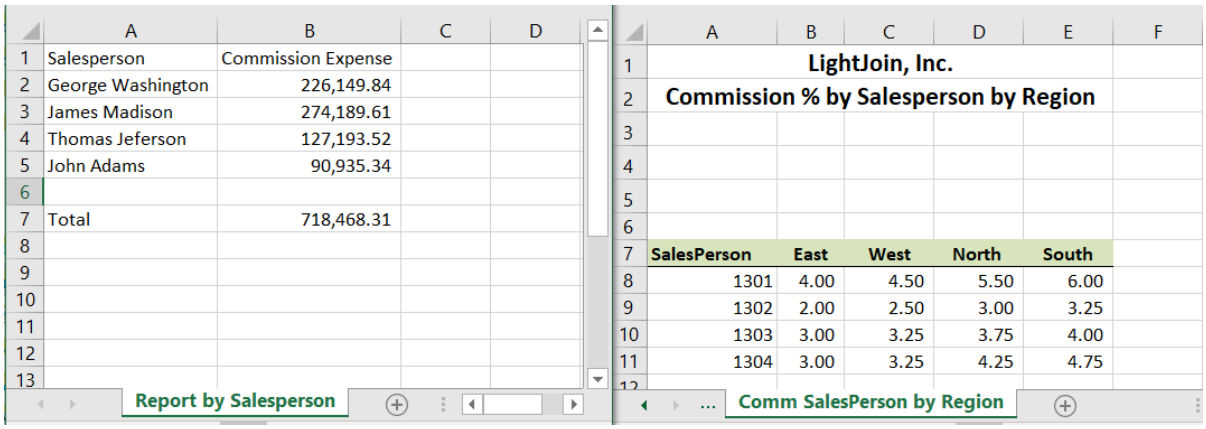
Report by Salesperson



	A	B	C
1	Salesperson	Commission Expense	
2	George Washington	226,149.84	
3	James Madison	274,189.61	
4	Thomas Jeferson	127,193.52	
5	John Adams	90,935.34	
6			
7	Total	718,468.31	
8			
9			
10			
11			
12			
13			

Report by Salesperson

Open the LightJoin.xlsx file and arrange both Excel files on the screen.



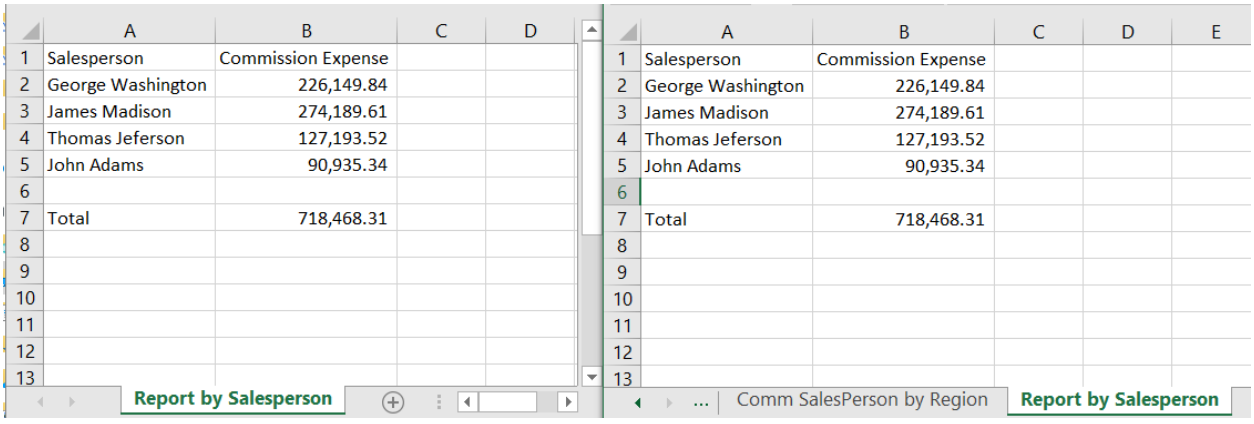
	A	B	C	D
1	Salesperson	Commission Expense		
2	George Washington	226,149.84		
3	James Madison	274,189.61		
4	Thomas Jeferson	127,193.52		
5	John Adams	90,935.34		
6				
7	Total	718,468.31		
8				
9				
10				
11				
12				
13				

Report by Salesperson

	A	B	C	D	E	F
1	LightJoin, Inc.					
2	Commission % by Salesperson by Region					
3						
4						
5						
6						
7	SalesPerson	East	West	North	South	
8	1301	4.00	4.50	5.50	6.00	
9	1302	2.00	2.50	3.00	3.25	
10	1303	3.00	3.25	3.75	4.00	
11	1304	3.00	3.25	4.25	4.75	
12						

... Comm SalesPerson by Region

Copy the “Report by Salesperson” sheet by left clicking on the title tab and holding down the Control key as you drag the file to the right of the “Comm SalesPerson by Region” sheet in the LongJoin file and releasing the mouse and then the Control key.



	A	B	C	D
1	Salesperson	Commission Expense		
2	George Washington	226,149.84		
3	James Madison	274,189.61		
4	Thomas Jeferson	127,193.52		
5	John Adams	90,935.34		
6				
7	Total	718,468.31		
8				
9				
10				
11				
12				
13				

Report by Salesperson

	A	B	C	D	E
1	Salesperson	Commission Expense			
2	George Washington	226,149.84			
3	James Madison	274,189.61			
4	Thomas Jeferson	127,193.52			
5	John Adams	90,935.34			
6					
7	Total	718,468.31			
8					
9					
10					
11					
12					
13					

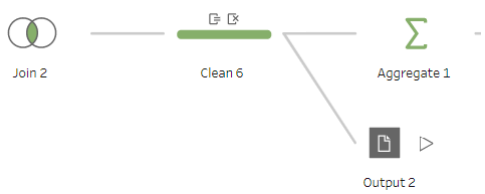
... Comm SalesPerson by Region | Report by Salesperson

Save and close both Excel files.

Step 16: Create a "Commission by Salesperson by Region" Report

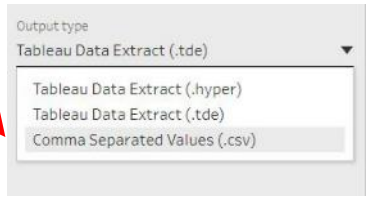
It is going to be much easier to create the "Commission by Salesperson by Region" report in Excel since all the data has already been transformed.

Add a step "Output 2" after the step "Clean 6".

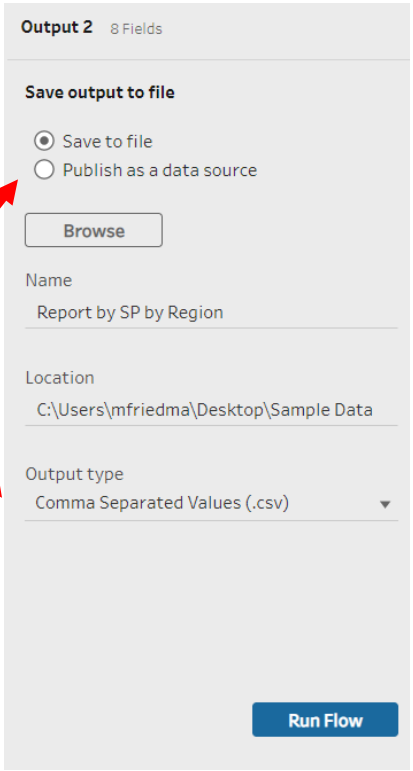


Verify that "Output type" is Comma Separated Values (.cvs)

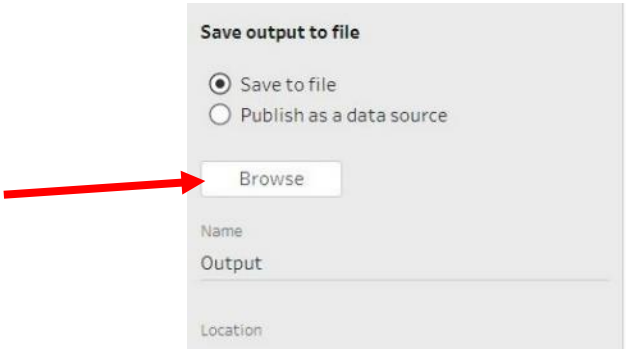
If not, then select the .csv file type.



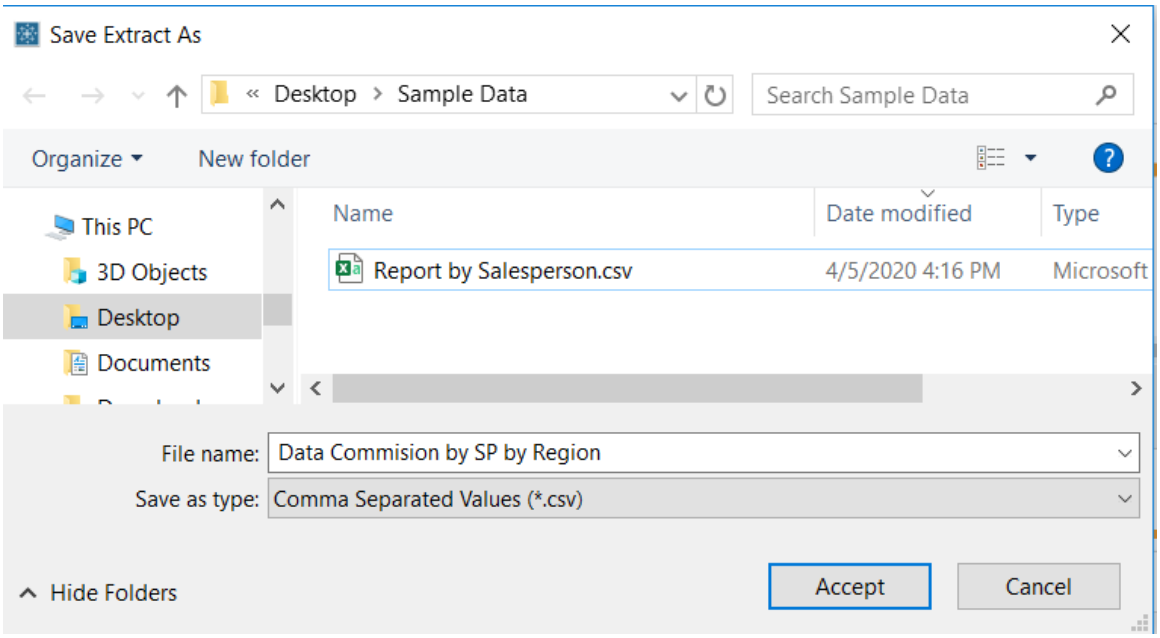
Select "Save to file" option button.



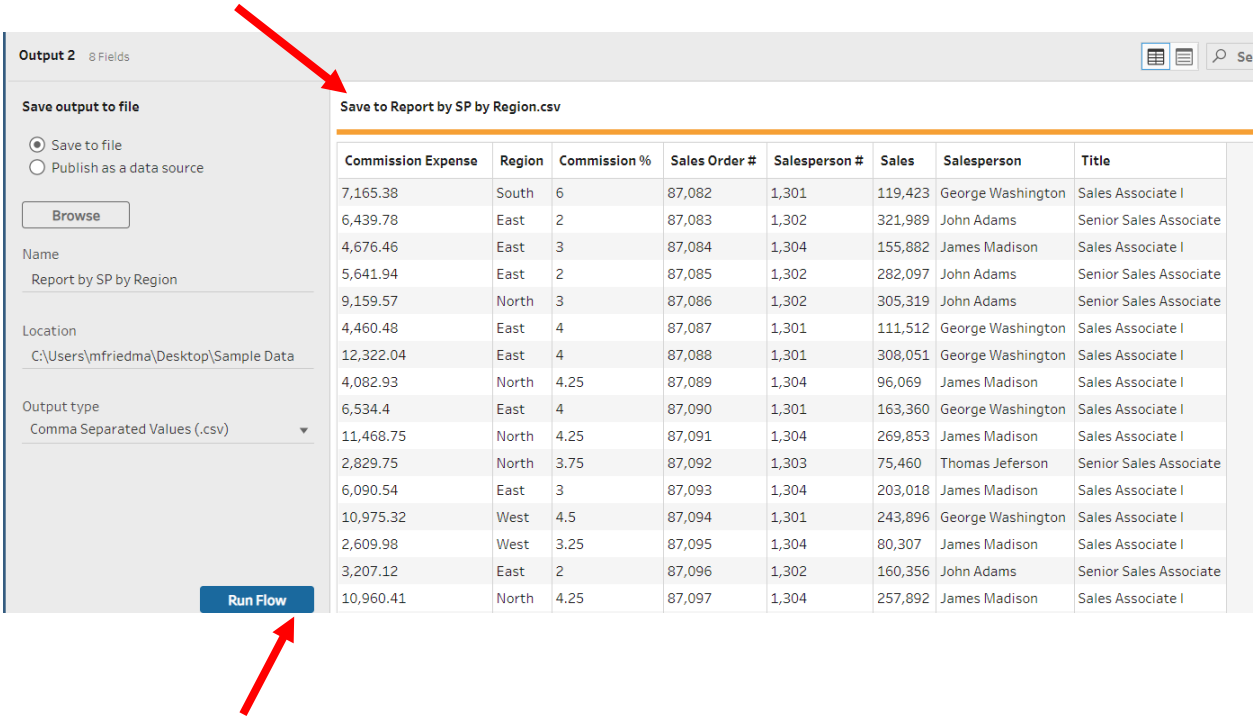
Select “Browse” button



Select the location for the file, specify the file name, ”Data Commission by SP and Region”, and select “Accept”.



The preview shows the data that will be generated.



Commission Expense	Region	Commission %	Sales Order #	Salesperson #	Sales	Salesperson	Title
7,165.38	South	6	87,082	1,301	119,423	George Washington	Sales Associate I
6,439.78	East	2	87,083	1,302	321,989	John Adams	Senior Sales Associate
4,676.46	East	3	87,084	1,304	155,882	James Madison	Sales Associate I
5,641.94	East	2	87,085	1,302	282,097	John Adams	Senior Sales Associate
9,159.57	North	3	87,086	1,302	305,319	John Adams	Senior Sales Associate
4,460.48	East	4	87,087	1,301	111,512	George Washington	Sales Associate I
12,322.04	East	4	87,088	1,301	308,051	George Washington	Sales Associate I
4,082.93	North	4.25	87,089	1,304	96,069	James Madison	Sales Associate I
6,534.4	East	4	87,090	1,301	163,360	George Washington	Sales Associate I
11,468.75	North	4.25	87,091	1,304	269,853	James Madison	Sales Associate I
2,829.75	North	3.75	87,092	1,303	75,460	Thomas Jeferson	Senior Sales Associate
6,090.54	East	3	87,093	1,304	203,018	James Madison	Sales Associate I
10,975.32	West	4.5	87,094	1,301	243,896	George Washington	Sales Associate I
2,609.98	West	3.25	87,095	1,304	80,307	James Madison	Sales Associate I
3,207.12	East	2	87,096	1,302	160,356	John Adams	Senior Sales Associate
10,960.41	North	4.25	87,097	1,304	257,892	James Madison	Sales Associate I

Select “Run Flow”

Select “Done”.

Navigate to the location and double click to open the CSV file you just saved.

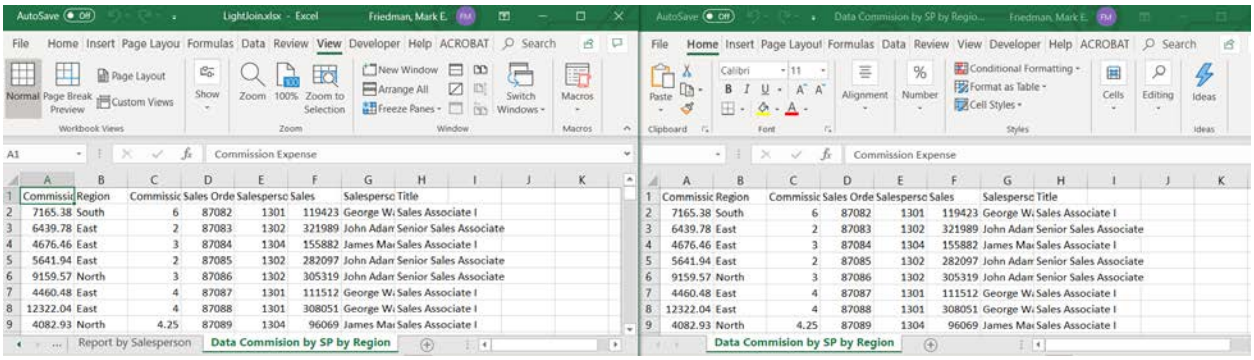
The exported table in Excel would like the following:

	A	B	C	D	E	F	G	H	I	J	K
1	Commissic	Region	Commissic	Sales Orde	Salesperso	Sales	Salesperso	Title			
2	7165.38	South	6	87082	1301	119423	George W	Sales Associate I			
3	6439.78	East	2	87083	1302	321989	John Adan	Senior Sales Associate			
4	4676.46	East	3	87084	1304	155882	James Ma	Sales Associate I			
5	5641.94	East	2	87085	1302	282097	John Adan	Senior Sales Associate			
6	9159.57	North	3	87086	1302	305319	John Adan	Senior Sales Associate			
7	4460.48	East	4	87087	1301	111512	George W	Sales Associate I			
8	12322.04	East	4	87088	1301	308051	George W	Sales Associate I			
9	4082.93	North	4.25	87089	1304	96069	James Ma	Sales Associate I			
10	6534.4	East	4	87090	1301	163360	George W	Sales Associate I			
11	11468.75	North	4.25	87091	1304	269853	James Ma	Sales Associate I			
12	2829.75	North	3.75	87092	1303	75460	Thomas Je	Senior Sales Associate			
13	6090.54	East	3	87093	1304	203018	James Ma	Sales Associate I			
14	10975.32	West	4.5	87094	1301	243896	George W	Sales Associate I			
15	2609.98	West	3.25	87095	1304	80307	James Ma	Sales Associate I			
16	3207.12	East	2	87096	1302	160356	John Adan	Senior Sales Associate			

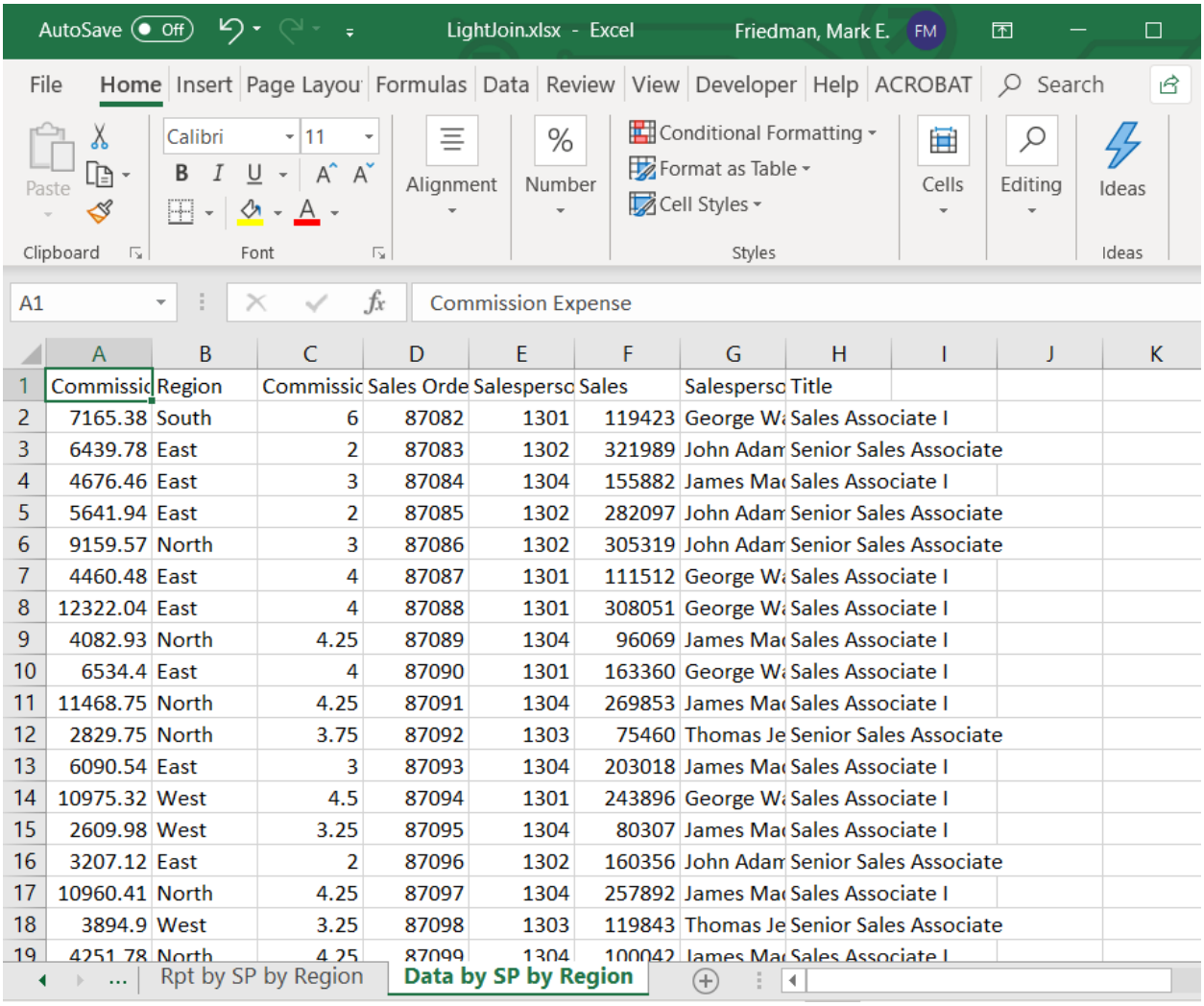
Data Commision by SP by Region

Open the LightJoin.xlsx file and arrange both Excel files on the screen.

Copy the “Data Commission by SP by Region” sheet from the .cvs file by left clicking on the title tab and hold down the Control key as you drag the file to the right of the “Report by Salesperson” sheet in the LongJoin.xlsx file and releasing the mouse and then the Control key.



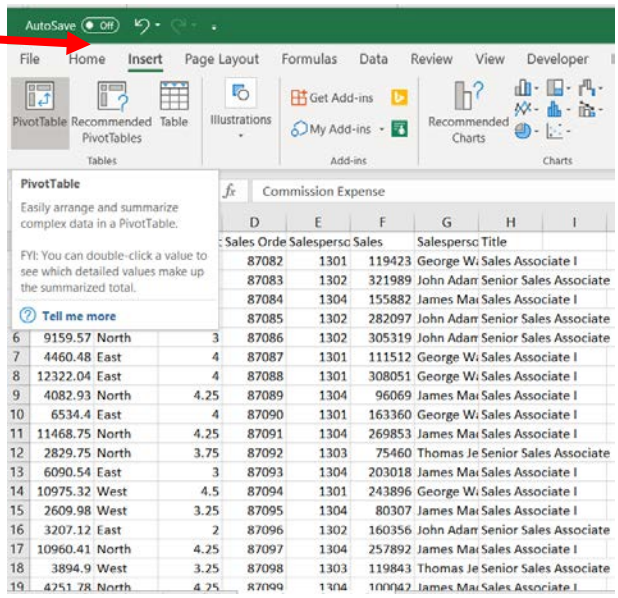
Select a cell, perhaps A1, in the data range on the “Data by SP by Region”.



	A	B	C	D	E	F	G	H	I	J	K
1	Commissic	Region	Commissic	Sales Orde	Salesperso	Sales	Salesperso	Title			
2	7165.38	South	6	87082	1301	119423	George W	Sales Associate I			
3	6439.78	East	2	87083	1302	321989	John Adam	Senior Sales Associate			
4	4676.46	East	3	87084	1304	155882	James Ma	Sales Associate I			
5	5641.94	East	2	87085	1302	282097	John Adam	Senior Sales Associate			
6	9159.57	North	3	87086	1302	305319	John Adam	Senior Sales Associate			
7	4460.48	East	4	87087	1301	111512	George W	Sales Associate I			
8	12322.04	East	4	87088	1301	308051	George W	Sales Associate I			
9	4082.93	North	4.25	87089	1304	96069	James Ma	Sales Associate I			
10	6534.4	East	4	87090	1301	163360	George W	Sales Associate I			
11	11468.75	North	4.25	87091	1304	269853	James Ma	Sales Associate I			
12	2829.75	North	3.75	87092	1303	75460	Thomas Je	Senior Sales Associate			
13	6090.54	East	3	87093	1304	203018	James Ma	Sales Associate I			
14	10975.32	West	4.5	87094	1301	243896	George W	Sales Associate I			
15	2609.98	West	3.25	87095	1304	80307	James Ma	Sales Associate I			
16	3207.12	East	2	87096	1302	160356	John Adam	Senior Sales Associate			
17	10960.41	North	4.25	87097	1304	257892	James Ma	Sales Associate I			
18	3894.9	West	3.25	87098	1303	119843	Thomas Je	Senior Sales Associate			
19	4251.78	North	4.25	87099	1304	100042	James Ma	Sales Associate I			

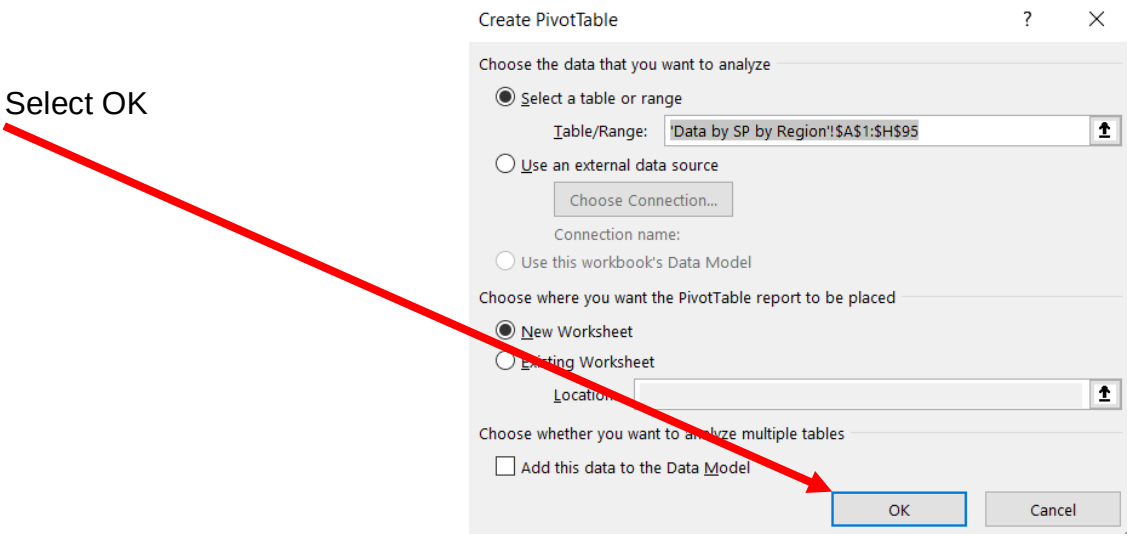
Select Insert

Select Pivot table



This will insert a Pivot Table in a new sheet. Pivot tables help summarize the data in various formats.

Select OK

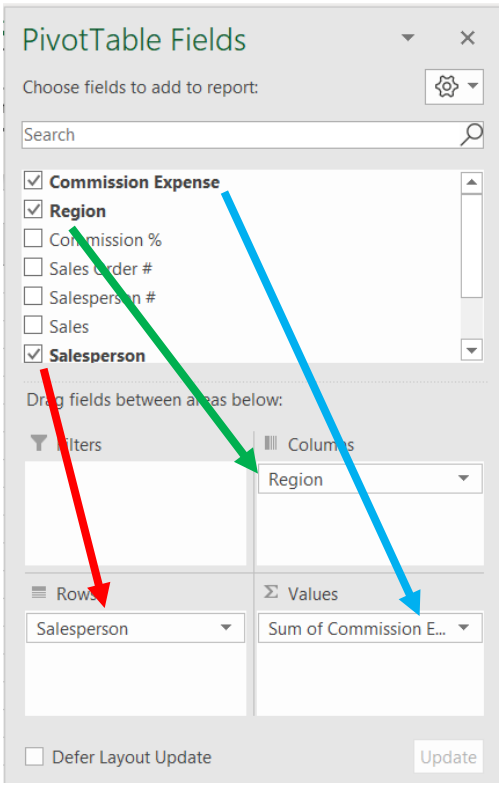


Once a Pivot table is inserted, the Pivot Table Fields are displayed .

Drag the fields

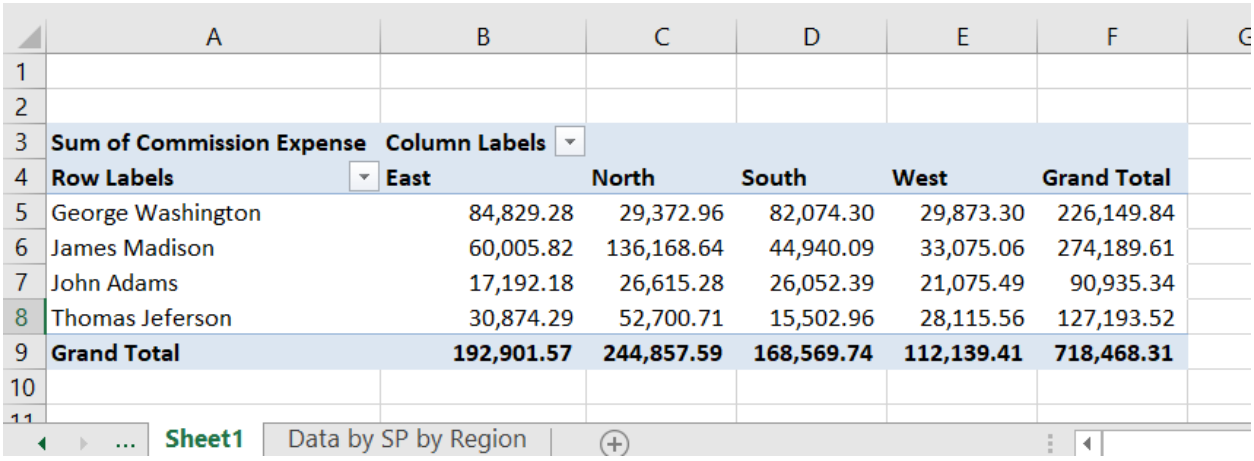
1. **Salesperson to Rows**
2. **Region to Columns** and
3. **Commission Expense to Values**

As indicated in the diagram



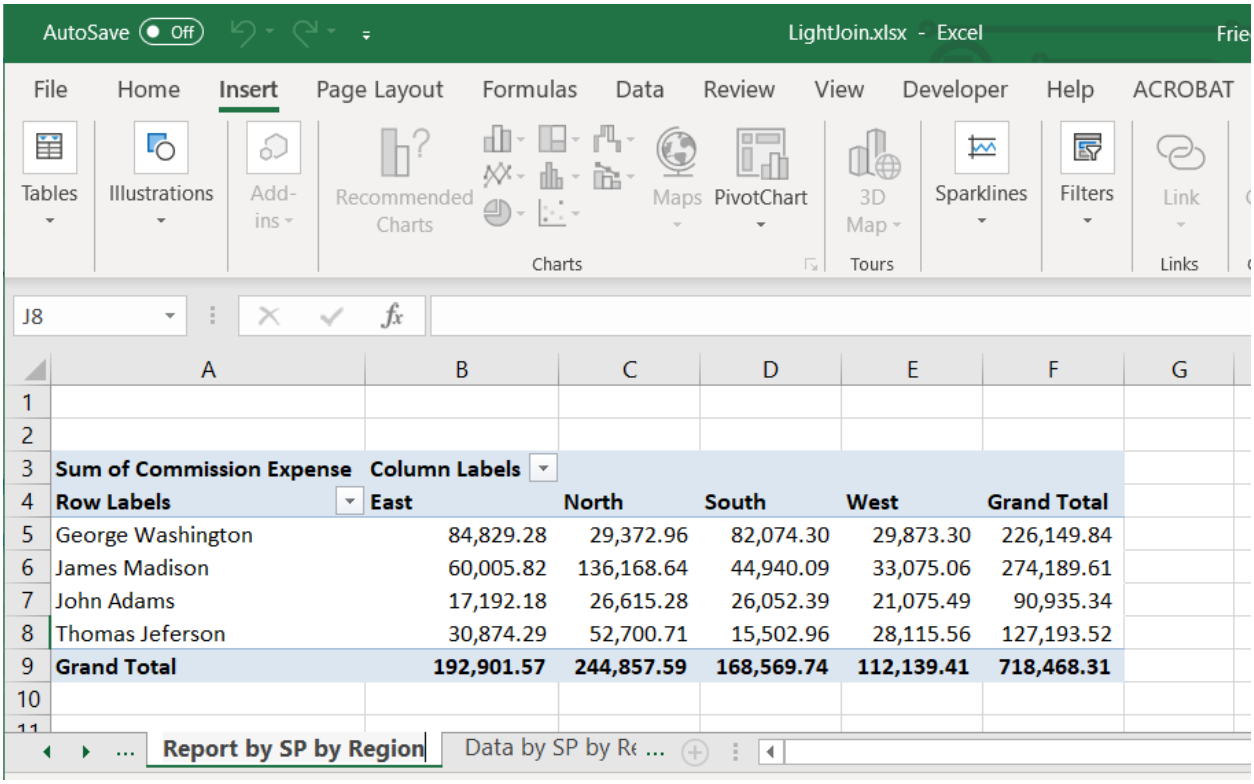
If “Count of Commission” appears instead of the “Sum of Commission”, left click on Count of Commissions, select Value.Field Setting, Sum then OK.

This summarizes the data in the following format:



Sum of Commission Expense	Column Labels				
Row Labels	East	North	South	West	Grand Total
George Washington	84,829.28	29,372.96	82,074.30	29,873.30	226,149.84
James Madison	60,005.82	136,168.64	44,940.09	33,075.06	274,189.61
John Adams	17,192.18	26,615.28	26,052.39	21,075.49	90,935.34
Thomas Jeferson	30,874.29	52,700.71	15,502.96	28,115.56	127,193.52
Grand Total	**192,901.57**	**244,857.59**	**168,569.74**	**112,139.41**	**718,468.31**

Rename “Sheet1” to “Report by SP by Region”



Sum of Commission Expense	Column Labels				
Row Labels	East	North	South	West	Grand Total
George Washington	84,829.28	29,372.96	82,074.30	29,873.30	226,149.84
James Madison	60,005.82	136,168.64	44,940.09	33,075.06	274,189.61
John Adams	17,192.18	26,615.28	26,052.39	21,075.49	90,935.34
Thomas Jeferson	30,874.29	52,700.71	15,502.96	28,115.56	127,193.52
Grand Total	**192,901.57**	**244,857.59**	**168,569.74**	**112,139.41**	**718,468.31**

Save and close the LongJoin.xls file.

Close the “Report by SP and Region.cvs” file.

Save and close the Tableau Prep flow file.

Done!