Business Intelligence Intermediate – Unit 3

Xccelerate - Data Science Immersive

Analyzing Data: Table Calculations

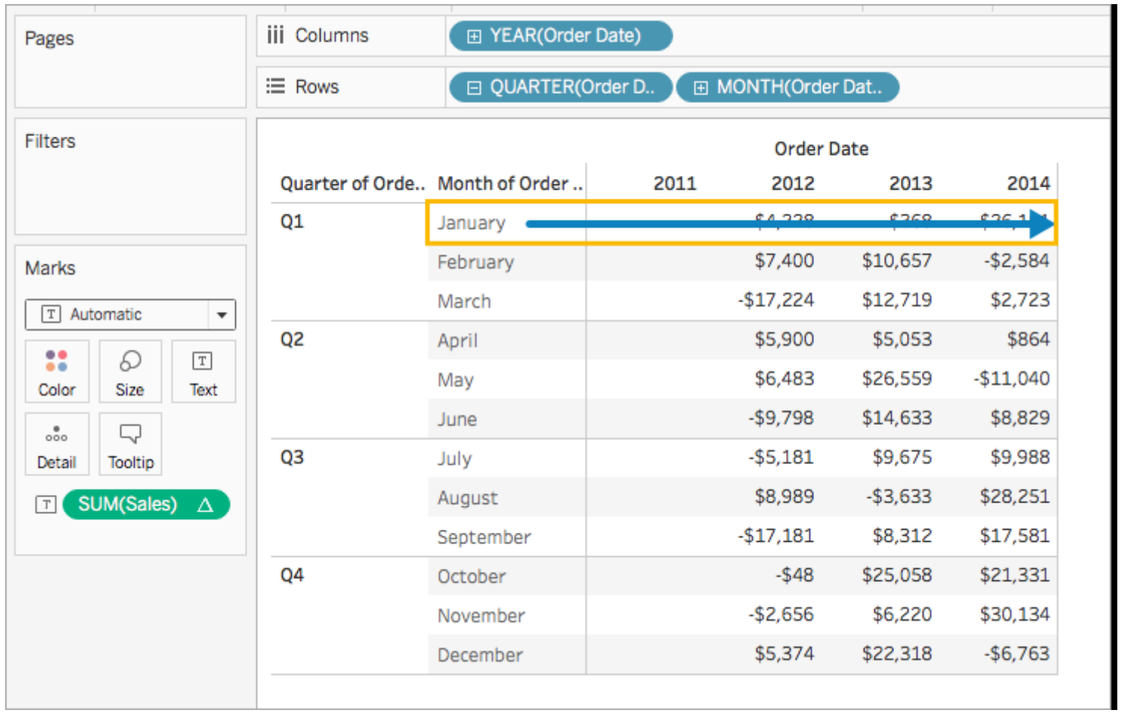
Transform Values with Table Calculations

- Table Calculations are based on the values in the view(fieldset) and not based on the values in the data source

- The dimensions that define how to group the calculation (the scope of data it is performed on) are called partitioning fields. The table calculation is performed separately within each partition.

- Table (across)
- Computes across the length of the table and restarts after every partition.

- For example, in the following table, the calculation is computed across columns (YEAR(Order Date)) for every row (MONTH(Order Date)).



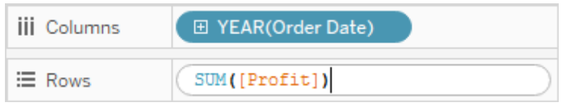
		Order Date			
Quarter of Orde..	Month of Order ..	2011	2012	2013	2014
Q1	January		[illegible]	[illegible]	[illegible]
	February		$7,400	$10,657	-$2,584
	March		-$17,224	$12,719	$2,723
Q2	April		$5,900	$5,053	$864
	May		$6,483	$26,559	-$11,040
	June		-$9,798	$14,633	$8,829
Q3	July		-$5,181	$9,675	$9,988
	August		$8,989	-$3,633	$28,251
	September		-$17,181	$8,312	$17,581
Q4	October		-$48	$25,058	$21,331
	November		-$2,656	$6,220	$30,134
	December		$5,374	$22,318	-$6,763

https://onlinehelp.tableau.com/current/pro/desktop/en-us/calculations_tablecalculations.htm

Analyzing Data: Ad-Hoc Calculation

Ad-Hoc Calculation

- Double-click on an existing field to start editing or just type in the empty shelf



- You can have multi-line Ad-Hoc Calculations and aggregated Ad-Hoc calculations

- If Tableau determines that the expression you enter is a measure (that is, returns a number), it automatically adds an aggregation to the expression when you commit the expression. For example, if you type DATEDIFF('day',[Ship Date],[Order Date]) in an ad-hoc calculation and then press Enter, what you will see is the following:

  SUM(DATEDIFF('day',[Ship Date],[Order Date]))

- If you use a field that is already an aggregated field (for example, SUM([Profit])) in an ad-hoc calculation, the result is an aggregate calculation. For example, when you commit an ad-hoc calculation SUM([Profit])/SUM([Sales]), the result is:

  AGG(SUM([Profit])/SUM([Sales]))

Analyzing Data: Sets

Sets: Sets are custom fields that define a subset of data based on some conditions.

Use sets in the visualization

After you create a set, it displays at the bottom of the Data pane in the Sets section. You can drag it into the viz like any other field.
When you drag a set to the viz in Tableau Desktop, you can choose to show the members of the set or aggregate the members into In/Out categories.

Show In/Out members in a set

In most cases, when you drag a set to the viz, Tableau displays the set using the In/Out mode. This mode separates the set into two categories:
In - The members in the set.
Out - Any members that are not part of the set.
For example, in a set defined for the top 25 customers, the top customers would be part of the In category and all other customers would be part of the Out category.

Analyzing Data: Sets

Step 1: Create the parameter
connect to the Sample-Superstore data source.
Name: Top Customers 2.
Data type: Integer.
Current value: 5.
Min: 5, Max:20, Step:5

Step 2: Create the top N customers set
Customer Name, and select Create > Set.
Name: Top N Customers by Sales.
Select By Field, From the field drop-down list (Category), select Sales.
From the aggregation drop-down list, select Sum.

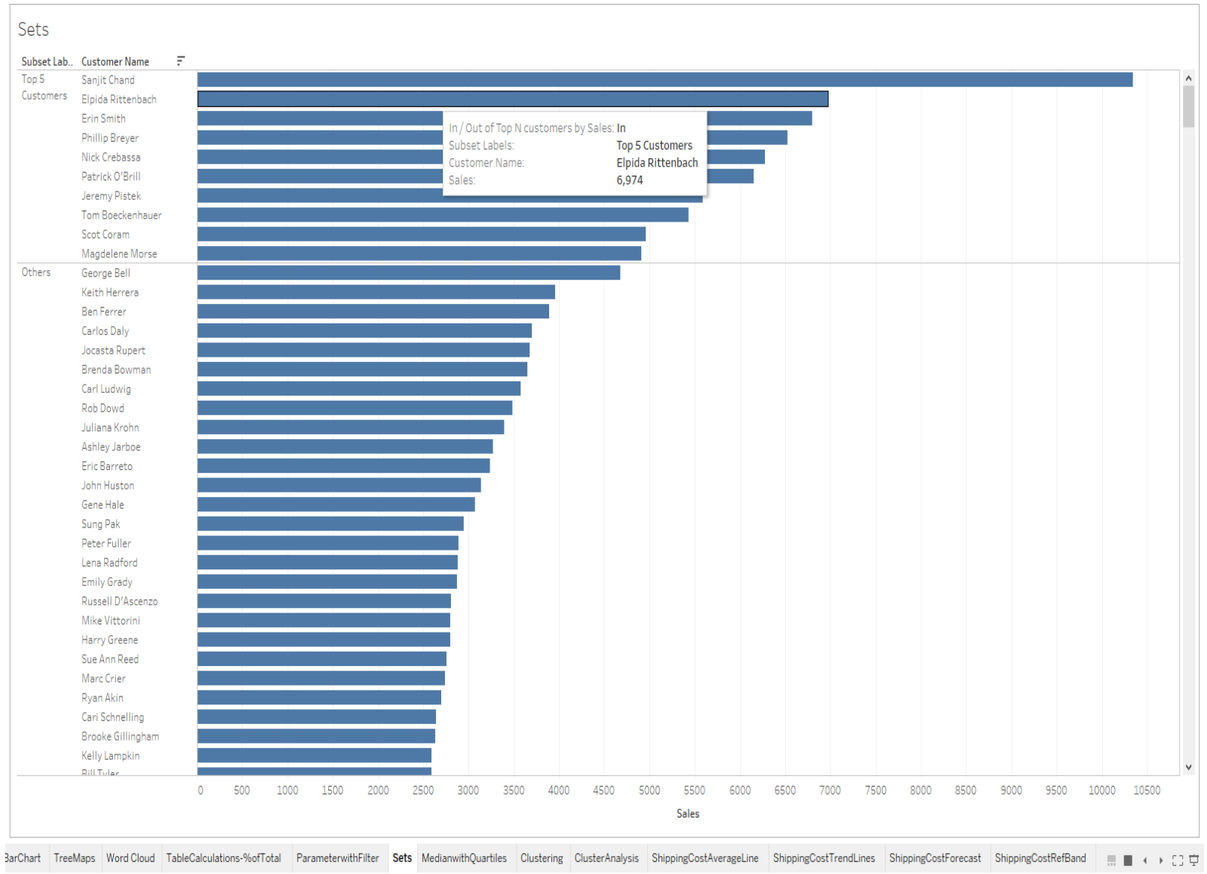
Step 3: Set up the view
From Sets, drag Top N Customers by Sales to the Rows shelf.
From Dimensions, drag Customer Name to the Rows shelf, positioning it to the right of the set.
From Measures, drag Sales to the Columns shelf.

On the toolbar, click the Sort Descending button to make sure that the set is working.

In the Data pane, under Sets, right-click Top N Customers by Sales, and then click Create Calculated Field.

In the Name text box, type Subset Labels.
In the Formula text box, type the following formula to create dynamic labels for the customers in the set:

```
IF [Top N Customers by Sales]
THEN "Top " + str([Top Customers 2]) + " Customers"
ELSE "Others"
END
```



Analyzing Data: Sets

From Dimensions, drag Subset Labels to the Rows shelf, placing it between the Top N set and the Customer Name dimension.

On the Rows shelf, right-click the IN/OUT(Top N Customers by Sales) set, and then clear Show Header.

This hides the In/Out labels while retaining the sort order so that your top N subset always appears at the top of the view.

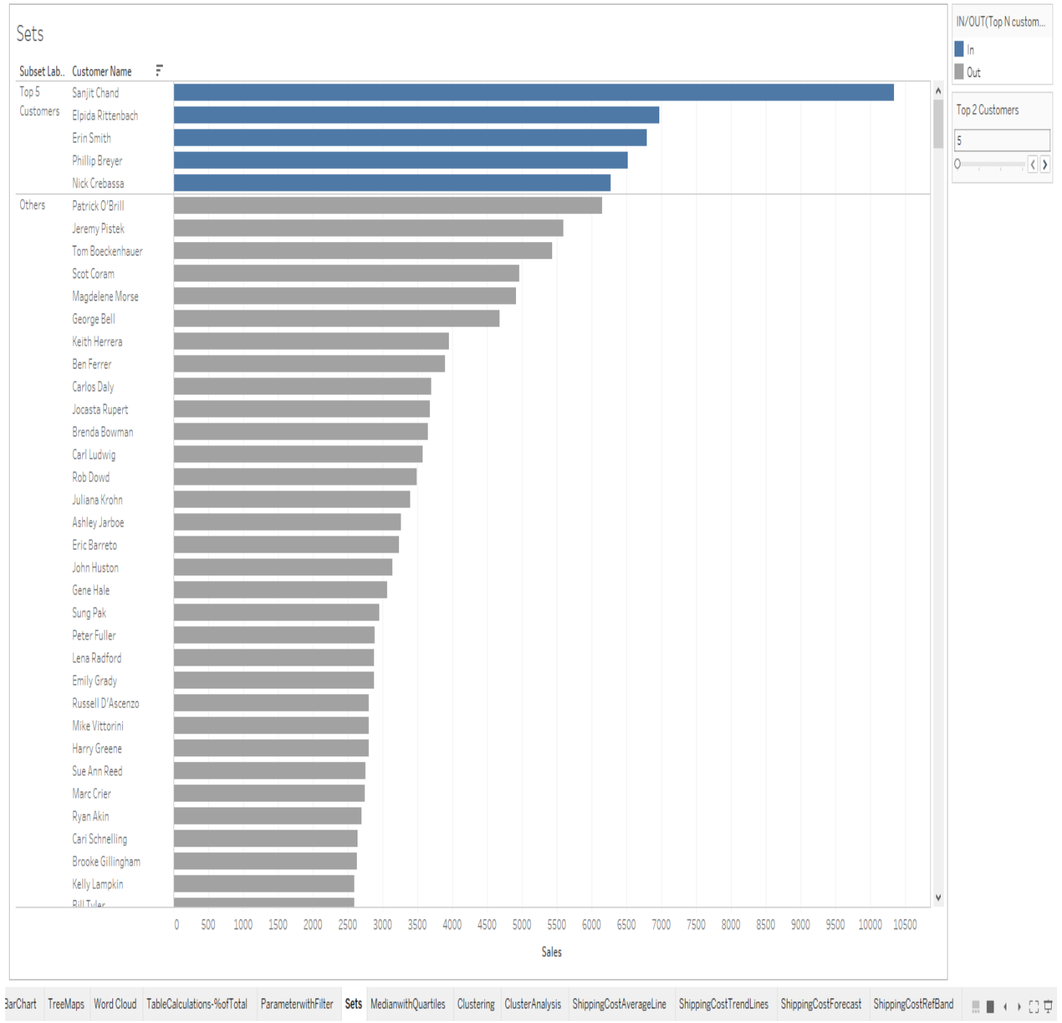
From Sets, drag Top N Customers by Sales to Color on the Marks card.

Step 4: Combine the Top N set with a dynamic parameter
In the Data pane, right-click Top N Customers by Sales, and then select Edit Set.
In the Edit Set dialog box, do the following:

Select the Top tab. Click the value drop-down menu, and select the Top Customers 2 parameter. This links the Top N Customers by Sales set to the Top Customers 2 dynamic parameter, instead of to a static list of 10.

This parameter will be used in combination with the Top N Customers by Sales set, to adjust the top N value in the view.

In the Data pane, under Parameters, right-click the Top Customers 2 parameter, and select Show Parameter Control. You can control the top N value by using the Top Customers 2 parameter control that appears in the view



https://onlinehelp.tableau.com/current/pro/desktop/en-us/sortgroup_sets_topn.htm

Agenda

TABLEAU:

- Brief Description of Advanced Analytics tools in Tableau
- Simple Use case Example
- Exercise: Clustering
- PRACTICE MAKES PERFECT PART3
- BONUS: Python Integration with Tableau: TabPy

UNIT 3:
INTRODUCTION TO ADVANCED ANALYTICS TOOLS

In this section, we will describe the different types of advanced analytics tools provided by Tableau

INTRODUCTION TO ADVANCED ANALYTICS TOOLS

- Once you are in the analytics pane, you will have access to a lot of different reference lines and other analytics tools.

- These can be applied to a whole table, each pane of the table or even to each cell, as well to dual axis charts.

- These can give simple averages or constant lines to the graphs, complex box plots , forecast future values, spot trends or cluster similar observations.

- There are a lot of options to choose from, so it's important to have in mind exactly what you want to show and only then find the option that suits your needs.

INTRODUCTION TO ADVANCED ANALYTICS TOOLS

BREAKDOWN OF THE ANALYTICS TAB

- **SUMMARIZE:**

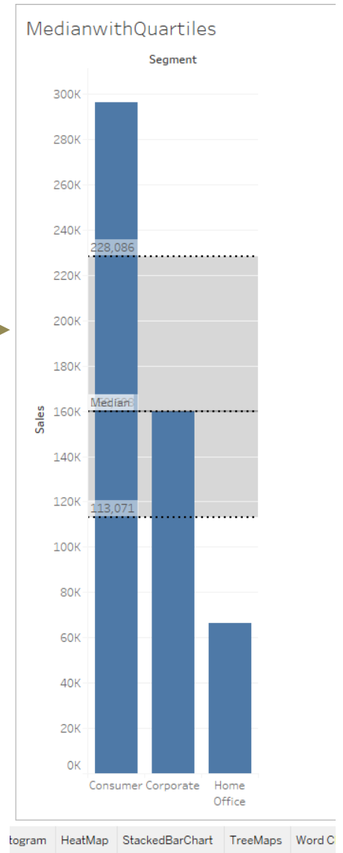
Constant Line - A line based on an inputted value.

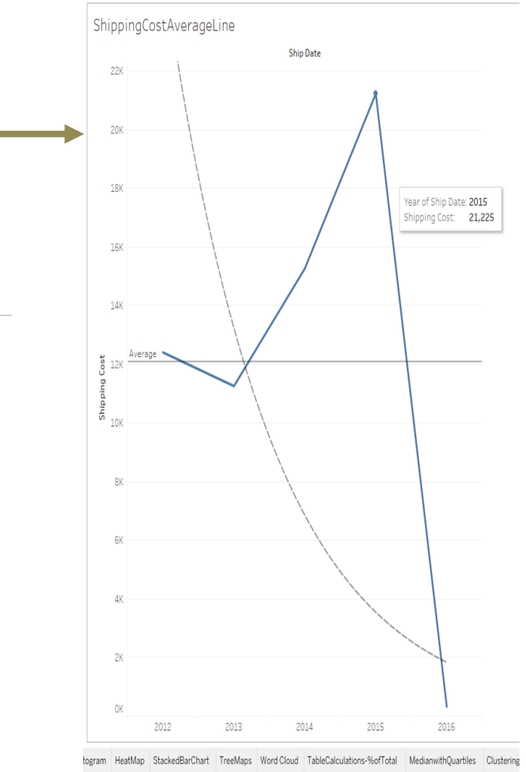
Average Line - A line showing the average summed up to the level specified.

Median with Quartiles - Median line and the lines which contain 25% and 75% of the values.

Box Plot – Add Box and Whisker plots to the view.

Totals - Adds totals to the view. When you add totals, the drop options are Subtotals, Column Grand Totals, and Row Grand Totals.





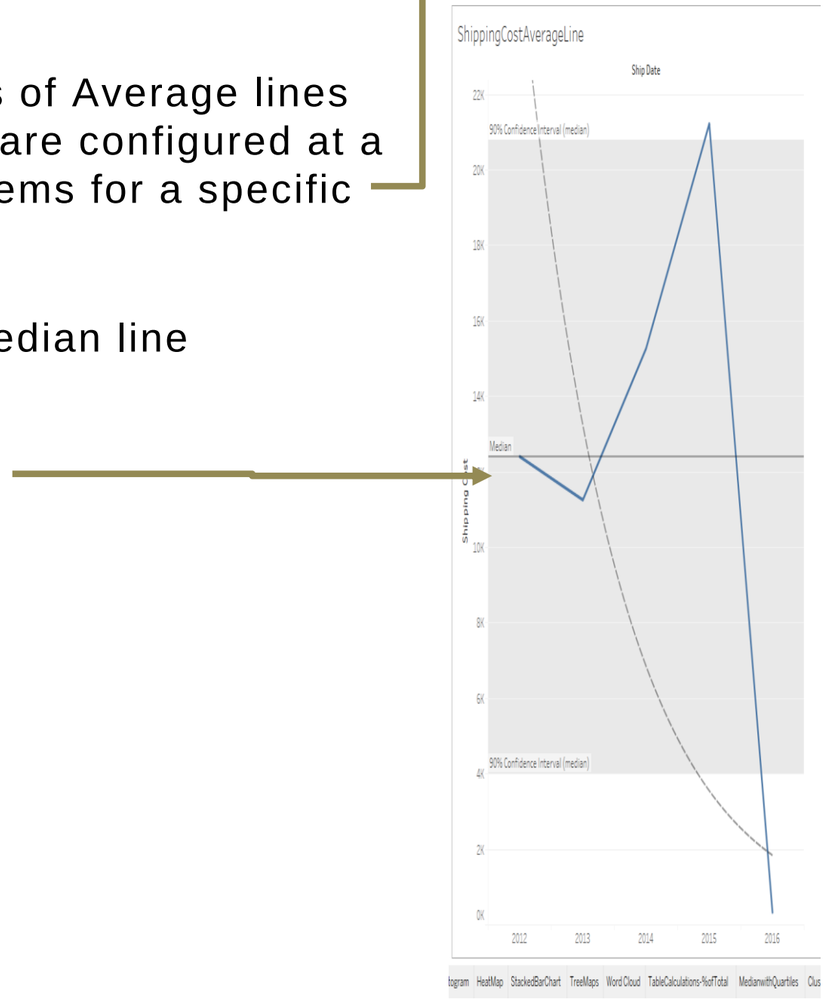
INTRODUCTION TO ADVANCED ANALYTICS TOOLS

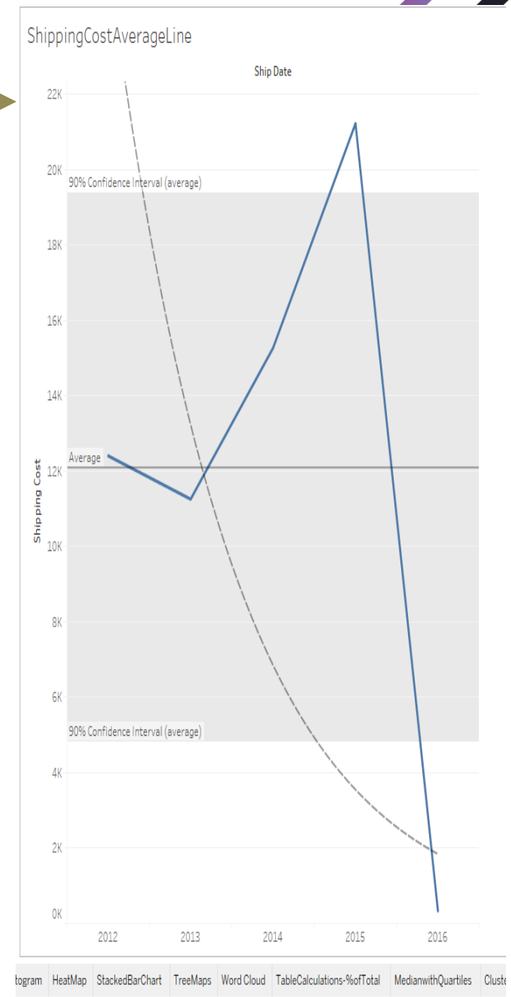
BREAKDOWN OF THE ANALYTICS TAB

- **MODEL:**

Average with 95% CI - Adds one or more sets of Average lines with distribution bands; the distribution bands are configured at a 95% confidence interval. You can add these items for a specific measure or for all measures.

Median with 95% CI – Same as above with Median line





Let’s understand confidence levels

Imagine you asked 50 customers how satisfied they were with their recent experience with your product on an 7 point scale, with 1 = not at all satisfied and 7 = extremely satisfied.

5 steps to calculate confidence levels

1. Find the mean by adding up the scores for each of the 50 customers and divide by the total number of responses (which is 50). For the purpose of this example, I have an average response of 6.

2. Compute the standard deviation. I have a sample standard deviation of 1.2.

3. Compute the standard error by dividing the standard deviation by the square root of the sample size: 1.2/ √(50) = .17.

4. Compute the margin of error by multiplying the standard error by 2. 17 x 2 = .34.

5. Compute the confidence interval by adding the margin of error to the mean from Step 1 and then subtracting the margin of error from the mean:

5.96+.34=6.3; 5.96-.34=5.6

We now have a 95% confidence interval of 5.6 to 6.3.

Our best estimate of what the entire customer population’s average satisfaction is between 5.6 to 6.3.

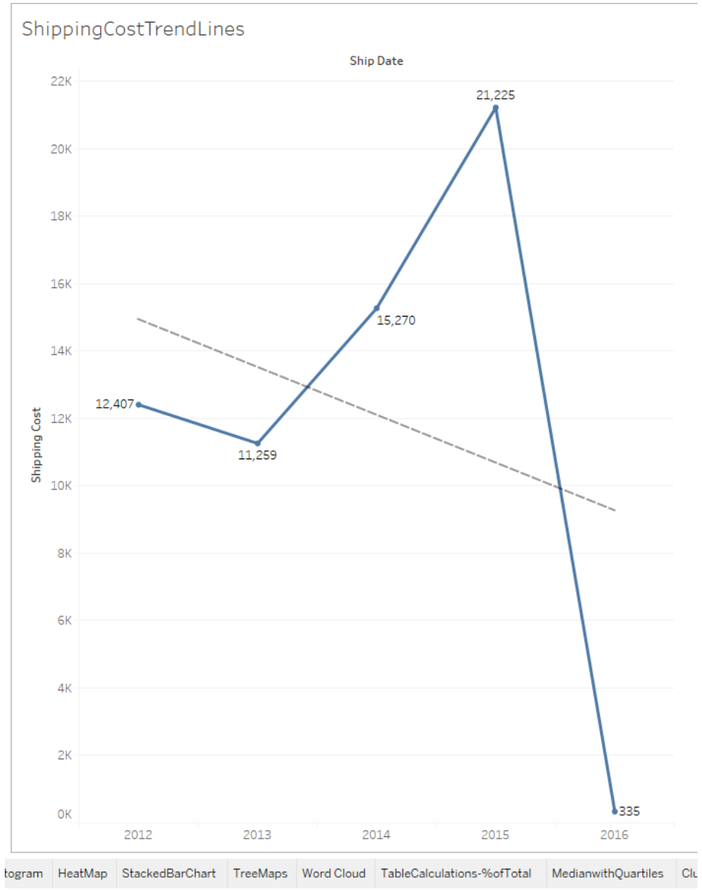
INTRODUCTION TO ADVANCED ANALYTICS TOOLS

BREAKDOWN OF THE ANALYTICS TAB

- **MODEL:**

Trend Line – You can show trend lines in a visualization to highlight trends in your data

Fit a trend line using one of the following regression model: Linear / Logarithmic / Exponential / Polynomial.



INTRODUCTION TO ADVANCED ANALYTICS TOOLS

BREAKDOWN OF THE ANALYTICS TAB

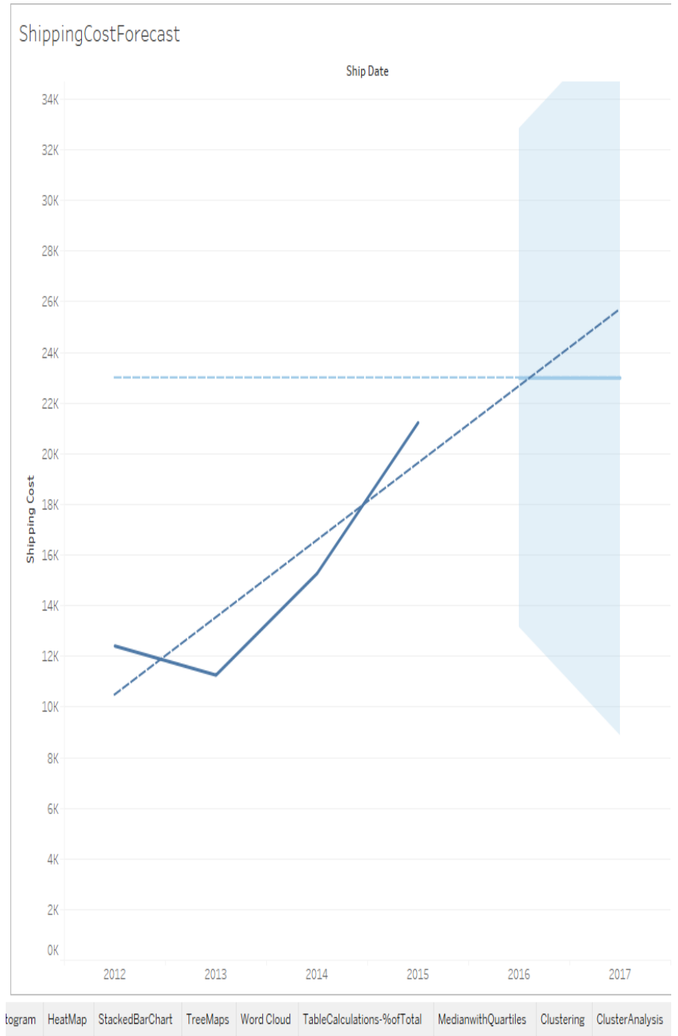
- **MODEL:**

Forecast – Forecasting in Tableau uses a technique known as Exponential smoothing.

Exponential smoothing models iteratively forecast future values of a regular time series of values from weighted averages of past values of the series.

The simplest model, Simple Exponential Smoothing, computes the next level or smoothed value from a weighted average of the last actual value and the last level value.

The method is exponential because the value of each level is influenced by every preceding actual value to an exponentially decreasing degree—more recent values are given greater weight.



INTRODUCTION TO ADVANCED ANALYTICS TOOLS

BREAKDOWN OF THE ANALYTICS TAB

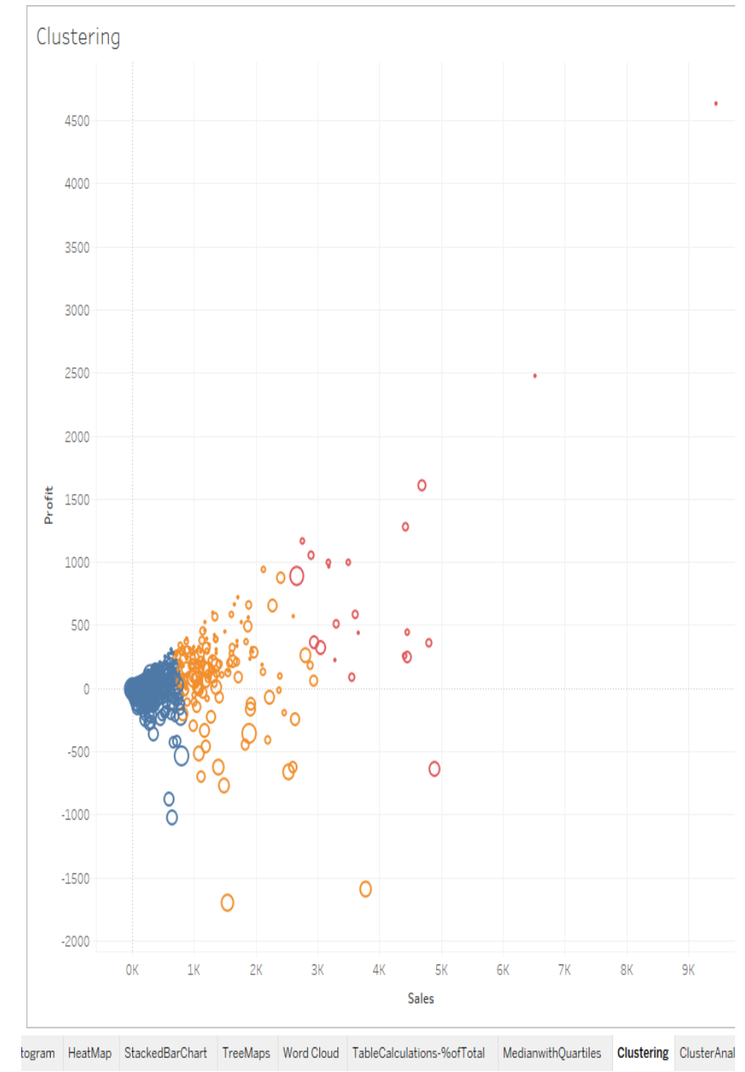
- **MODEL:**

Clustering – Tableau uses the k-means algorithm for clustering. you can specify a desired number of clusters, or have Tableau test different values of k and suggest an optimal number of clusters.

Tableau uses the Calinski-Harabasz criterion to assess cluster quality.

ClusterAnalysis

3Clusters	Discount	Distint..	Profit	Sales
Cluster 1	283	560	20,315	214,063
Cluster 2	38	228	21,207	218,913
Cluster 3	5	51	19,240	89,640



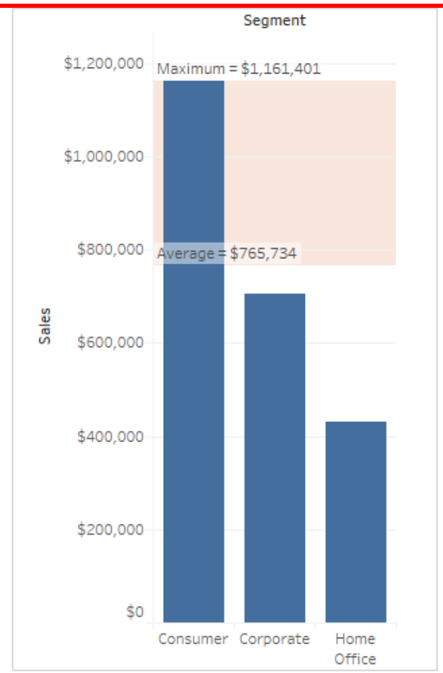
INTRODUCTION TO ADVANCED ANALYTICS TOOLS

BREAKDOWN OF THE ANALYTICS TAB

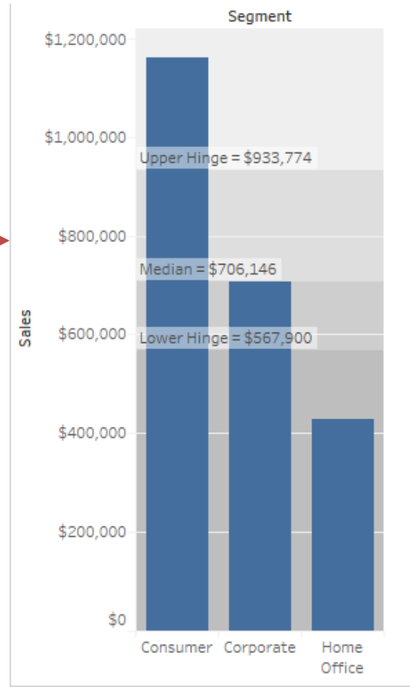
- **CUSTOM**

Reference Line - add a reference line at a constant or computed value on the axis. Computed values can be based on a specified field. You can also include confidence intervals with a reference line.

Reference Band - Reference bands shade an area behind the marks in the view between two constant or computed values on the axis.



Reference Distributions - Reference distributions add a gradient of shading to indicate the distribution of values along the axis. Distribution can be defined by percentages, percentiles, quantiles
(as in the following image), or standard deviation.



Box Plot – Box plot that can be customized.

INTRODUCTION TO ADVANCED ANALYTICS TOOLS

BREAKDOWN OF THE ANALYTICS TAB

Please refer to the online documentation for more information:

ANALYTICS PANE OVERVIEW:
https://onlinehelp.tableau.com/current/pro/desktop/en-us/environ_workspace_analytics_pane.htm

MODEL:
Trend Lines:
https://onlinehelp.tableau.com/current/pro/desktop/en-us/trendlines_add.htm#modeltype

Forecast:
https://onlinehelp.tableau.com/current/pro/desktop/en-us/forecasting.htm

Clustering:
https://onlinehelp.tableau.com/current/pro/desktop/en-us/clustering.htm

CUSTOM:
https://onlinehelp.tableau.com/current/pro/desktop/en-us/reference_lines.htm

ANALYSE DATA:
https://onlinehelp.tableau.com/current/pro/desktop/en-us/analyze.htm

UNIT 3:

SIMPLE USE CASE EXAMPLE

Clustering

SIMPLE USE CASE

Clustering is a powerful new feature in Tableau 10 that allows you to easily group similar dimension members.

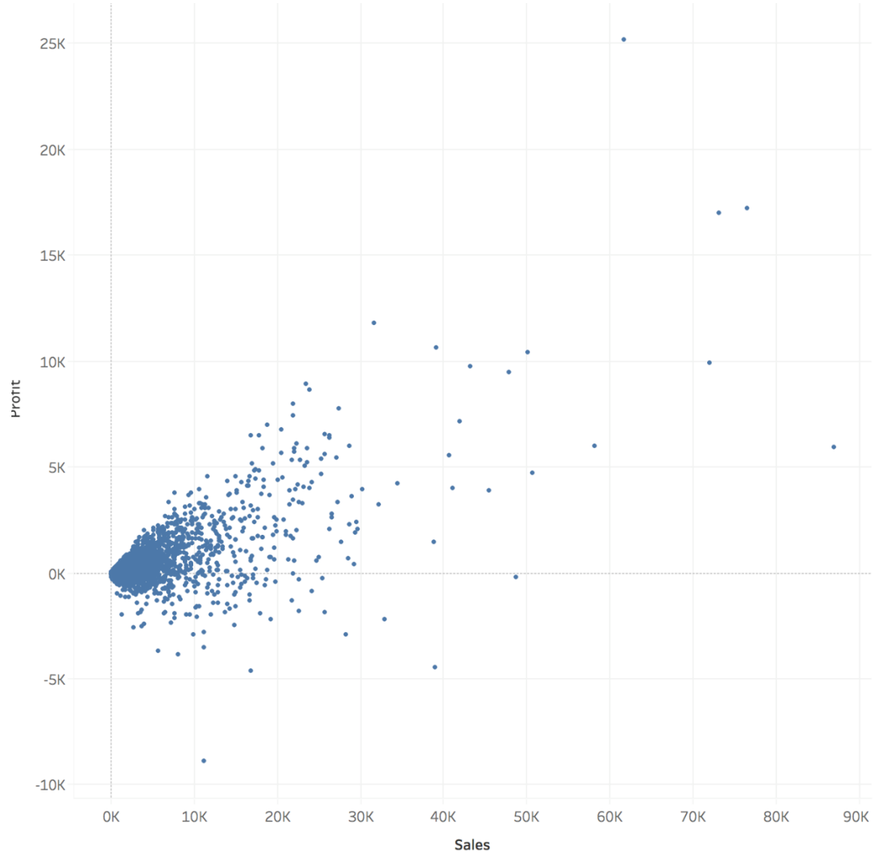
This type of clustering helps us create statistically-based segments which provide insight into how different groups are similar as well as how they are performing compared to each other.

Lets try to see how the clustering feature works using the sample **SuperStore dataset**.

To demonstrate cluster analysis lets start by building a scatter plot.

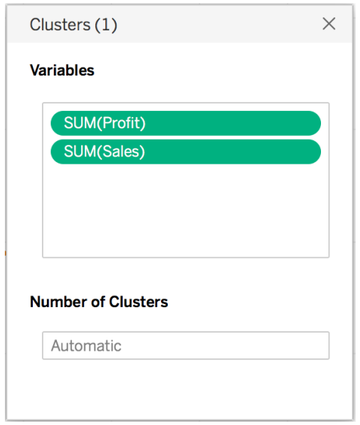
1) Bring Sales to the Columns shelf and Profit to the rows shelf.
1) Drag product name to the details shelf.

Your screen will now look like this:



SIMPLE USE CASE

Once we have the scatter plot ready lets go to the analytic pane and drag Cluster under models to our worksheet. As soon as we drop this onto the worksheet we can see a pop-up screen as follow :

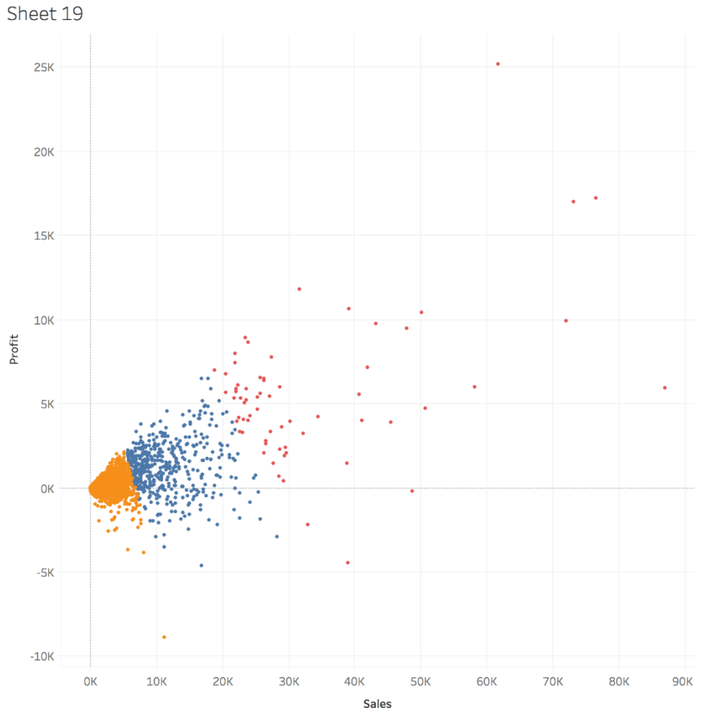


This pop-up box allows us to select the variables on which we want to create the clusters and define the number of clusters that we want in our dataset.

By default we can see only Profit and Sales variable in the box, we can add more variable by dragging it into the Variables space in the pop up window.

The number of clusters can be specified into the Number of Clusters space. Lets enter 3 in that space and close the pop-up window.

Our screen will now look as :



SIMPLE USE CASE

Here we can see that our scatter plot in now divided into three clusters- each represented by a different color.
There's much more we can do with the clusters that we created.

To edit the clusters right click on the Clusters in the colors mark and select Edit .

Describe clusters provides us with a summary statistics related to cluster analysis.

Following is the result when we select Describe Cluster from the dropdown.

Inputs for Clustering

Variables:	Sum of Profit
	Sum of Sales
Level of Detail:	Product Name
Scaling:	Normalized

Summary Diagnostics

Number of Clusters:	3
Number of Points:	3788
Between-group Sum of Squares:	15.222
Within-group Sum of Squares:	6.8291
Total Sum of Squares:	22.052

		Centers	
Clusters	**Number of Items**	**Sum of Profit**	**Sum of Sales**
Cluster 1	496	1253.6	11194.0
Cluster 2	3226	147.68	1520.4
Cluster 3	66	5594.7	33110.0
Not Clustered	0		

SIMPLE USE CASE

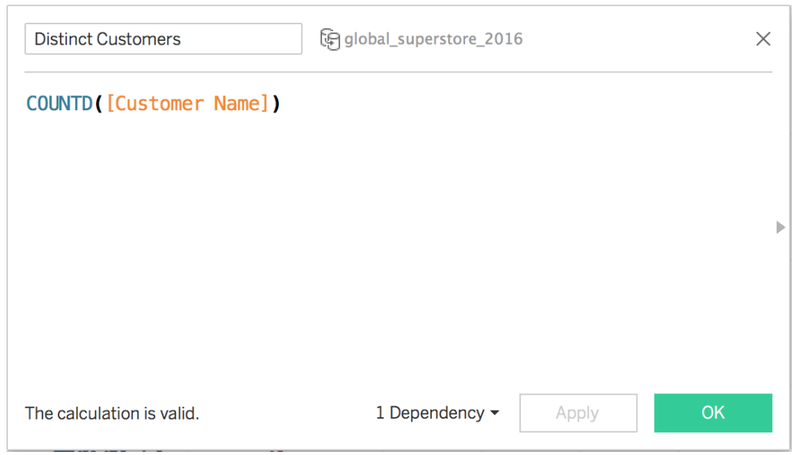
What next?

Lets say we want to see how many customers are in each cluster and what is the average sales, profit and discount for each of these clusters.

To do this, let’s start by dragging the Cluster variable from the color mark to the Dimensions area on the left , you will notice that a dimension for cluster gets created. Rename it as Cluster 3 (to signify that it has three clusters) .

Now, Open a new worksheet that connects to the same data source.

Create a calculated field to count the number of distinct customers:



SIMPLE USE CASE

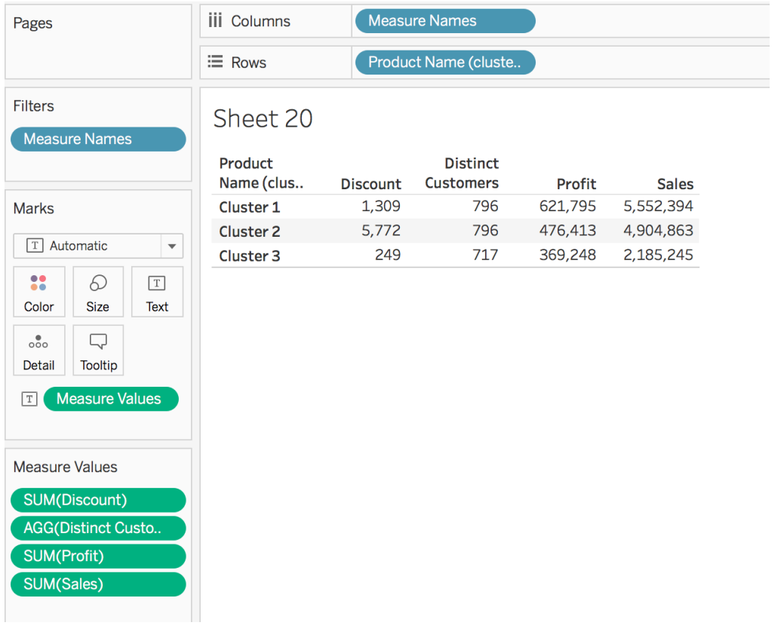
Drag the cluster dimension Cluster 3 to Rows.

Bring Measure names to the filter marks and select the measures Discount, Profit, Sales and Distinct Customers.

Drag measure names to the Column shelf and place measure values on the label mark.

Select the default aggregation of the measures (Profit, Sales and Discount) as average.

We will have the following table on our worksheet now:



Product Name (clus..	Discount	Distinct Customers	Profit	Sales
Cluster 1	1,309	796	621,795	5,552,394
Cluster 2	5,772	796	476,413	4,904,863
Cluster 3	249	717	369,248	2,185,245

In this table, we can easily see how the profit , sales and discount varies for each of the cluster created along with the number of customers that are under each clusters.

UNIT 3:

ADVANCED ANALYTICS IN TABLEAU

WALKTHROUGH EXERCISE

→ CLUSTERING

PRACTISE MAKES PERFECT

EXAM PREP QUESTIONS

PRACTISE MAKES PERFECT - Analytics

Exercise 1: Clustering (BI_Intermediate_Unit3_Exercise_Clustering.xlsx)

The Challenge: **Start Up Expansion Analytic**

You are a Data Analyst working for a tailored food delivery startup *WeCareUrHealth*. This is a relatively small company and they cannot compete with the big players in major cities.
They cook and deliver tailored-organic meal to their customers.
The company's strategy is to build a vast network in the smaller cities.

WeCareUrHealth already had a strong presence in 140 locations and recently opened stores in 10 new cities. Additionally, the company has two separate sales region.

1)
Identify which of the two sales regions is performing better (i.e. outperforms the other in 2 of the following 3 metrics):

- Average Revenue per Region
- Average Marketing spend per Region (less is better)
- Average ROMI (Return on Marketing Investment) per Region (Revenue/Marketing spend)

2)
Identify which of the 10 new locations have the best potential for the company to invest more funds into marketing.

PRACTISE MAKES PERFECT - Analytics

Exercise2: BI_Intermediate_Unit3_Exercise_Clustering.xlsx and BI_Intermediate_Unit3_Exercise_US_Population.csv

Instruction PART 1:

1) Create a MAP of U.S. Create custom territory to allow you to visualize the 3 necessary metric (Average Revenue per Region, Average Marketing spend and Average ROMI) aggregated at the Region level.
    Hint: you will have to group the cities belonging to Region 1 together.

Identify which one of your 2 sales region is performing better.

PRACTISE MAKES PERFECT - Analytics

Instruction PART 2:

2) Create a second Map of US:
→ Visualize in one Map the revenue generated by your old and new expansion cities.

3) Create a scatterplot of Revenue versus Marketing Spent.

Hint: We should have 2 distinct clusters in our visualization (Tableau Clustering Model tool).

Identify which of the 10 new locations have the best potential for the company to invest more funds into marketing ?

PRACTISE MAKES PERFECT - Clustering

For this exercise, you have access to 2 datasets:
BI_Intermediate_Unit3_Exercise_Clustering.xlsx and BI_Intermediate_Unit3_Exercise_US_Population.csv

Instruction PART 3:

Let’s add a bit of business knowledge and refine our clustering model.
We found previously 2 clusters and were able to identify the locations that have the best potential for the company.

However, we have access to the population demographic of all cities in US.
We know that our company would potentially perform better in cities with higher population proportion, in other worlds, the number of people in a city is directly correlated to the performance of our company.
In that sense, we can refine our Clustering model by bringing in this new information.

1) Perform a cross-database join on BI_Intermediate_Unit3_Exercise_Clustering.xlsx and BI_Intermediate_Unit3_Exercise_US_Population.csv
→ Hint: join the two files on multiple fields: Cities and State.

PRACTISE MAKES PERFECT - Clustering

For this exercise, you have access to 2 datasets:
BI_Intermediate_Unit3_Exercise_Clustering.xlsx and BI_Intermediate_Unit3_Exercise_US_Population.csv

Instruction PART 4:

2) Duplicate your previously created Cluster worksheet.
→ Then, left click into the Cluster mark, select edit clusters, drag and drop your Estimates2015 measure into the newly opened popup window, under your 2 other variables.

3) Analyze your newly created clusters.
→ Add a trend line for each of your clusters.

Identify which of the 10 new locations have the best potential for the company to invest more funds into marketing ?

PRACTICE MAKES PERFECT

PART 3 – Exam Prep

Please use the following dataset for all the questions: Certification_Prep_Dataset_Unit3.xlsx

Practice Makes Perfect

Please use the following dataset for all the questions:
Certification_Prep_Dataset_Unit3.xlsx

Question 1:

To connect to multiple tables in a single data source at one time, what must be specified?
a. A blend
b. A calculation
c. A join
d. A hierarchy

Question 2:

Tableau can create worksheet-specific filters.
a. True
b. False

*The following questions are practice questions for Tableau certification exams

Practice Makes Perfect

Please use the following dataset for all the questions:
Dataset_Unit3.xlsx

Question 3:

What does the box in a box plot represent?
a. Maximum extent of the data
b. The range of the middle half of the data points
c. The median of the middle half of the data points
d. The outliers of the data

Question 4:

What is the percent of total Sales for the ‘Home Office’ Customer Segment in July of 2012?
a. 23.50%
b. 23.97%
c. 20.14%
d. 32.56%

*The following questions are practice questions for Tableau certification exams

Practice Makes Perfect

Please use the following dataset for all the questions:
Dataset_Unit3.xlsx

Question 5:

Find the top 10 Product Names by Sales within each region. Which product is ranked #2 in both the Central & West regions in 2011?

a. Riverside Palais Royal Lawyers Bookcase
b. Bush Mission Pointe Library
c. Sharp AL-1530CS Digital Copier
d. Global Troy Executive Leather Low Back Tilter

Question 6:

In the Technology Product Category, which unprofitable state is surrounded by only profitable states?

a. Colorado
b. Missouri
c. Wyoming
d. Utah

*The following questions are practice questions for Tableau certification exams

Practice Makes Perfect

Please use the following dataset for all the questions:
Dataset_Unit3.xlsx

Question 7:

If 2013 Sales numbers were expected to increase by 50% in the following year, what would be the total estimated sales for the Consumer Segment in 2014?

a. $4,278,540
b. $816,999
c. $2,752,823
d. $802,365

Question 8:

In which Region do all Product Categories fall beneath the overall average profit?

a. All Regions
b. Central
c. East
d. South
e. West

*The following questions are practice questions for Tableau certification exams

Practice Makes Perfect

Please use the following dataset for all the questions:
Dataset_Unit3.xlsx

Question 9:

Which Product Sub-Category has a Shipping Cost to Sales ratio of above 3%?
a. Tables
b. Chairs & Chairmats
c. Paper
d. Binders and Binder Accessories

Question 10:

Find the customer with the lowest overall profit. What is his/her profit ratio?
a. 2.35%
b. 1%
c. -17.54%
d. -771.39%

*The following questions are practice questions for Tableau certification exams

Practice Makes Perfect

Please use the following dataset for all the questions:
Dataset_Unit4.xlsx

Question 11:

Determine which State in the Central Region has the highest distribution of profits using interquartile ranges.
a. South Dakota
b. North Dakota
c. Minnesota
d. Iowa

Question 12:

Look at the sum of profits for each Product Sub-Category. Which sub-category is $31,069 below the average profit across all categories?
a. Appliances
b. Bookcases
c. Envelopes
d. Paper

*The following questions are practice questions for Tableau certification exams

Practice Makes Perfect

Question 13:

What percent of total profits do the top 10 customer by Sales represent?
a. 3.50%
b. 5.03%
c. 17.54%
d. None of the Above

Question 14:

What was the Moving Average of Sales in June of 2012, including six months prior and six months after?
a. $101,752
b. $180,036
c. $188,552
d. $286,170

Hint: use table calculation

*The following questions are practice questions for Tableau certification exams

Practice Makes Perfect

Question 15:

Create a histogram showing the number of Sales using Sales Bins of $1,000. Which bins have profit ratios
(profit as a percentage of sales) of more than 25%? (Select all that apply)

a. 1,000
b. 3,000
c. 7,000
d. 8,000
e. 10,000
f. 11,000
g. 18,000

Kindly note:
Solutions to those exercises will be given to you at the end of the day

*The following questions are practice questions for Tableau certification exams

UNIT 3:

BONUS

BONUS

TABLEAU & PYTHON:

You can now integrate your Python code into Tableau with the library name TabPy.

When you use TabPy with Tableau, you can define calculated fields in Python, thereby leveraging the power of a large number of machine-learning libraries right from your visualizations.

Read more at https://www.tableau.com/about/blog/2016/11/leverage-power-python-tableau-tabpy-62077#mSGHwD5R5bSTBOry.99

Kindly note that The TabPy Library is till at an early stage of development, hence the integration is not bullet proof and TabPy can **only integrate with Tableau Desktop** and Server.

For more information, please visit the link below:
https://github.com/tableau/TabPy