Course Description  
In this course you WILL experience firsthand all of the PAIN a Data Scientist goes through on a daily basis. Corrupt data, anomalies, irregularities - you name it!

This course will give you a full overview of the Data Science journey. Upon completing this course you will know:

* How to clean and prepare your data for analysis
* How to perform basic visualisation of your data
* How to model your data
* How to curve-fit your data
* And finally, how to present your findings and wow the audience

This course will give you so much practical exercises that real world will seem like a piece of cake when you graduate this class. This course has homework exercises that are so thought provoking and challenging that you will want to cry... But you won't give up! You will crush it. In this course you will develop a good understanding of the following tools:

* SQL
* SSIS
* Tableau
* Gretl

**What you’ll learn**

* Successfully perform all steps in a complex Data Science project
* Create Basic Tableau Visualisations
* Perform Data Mining in Tableau
* Understand how to apply the Chi-Squared statistical test
* Apply Ordinary Least Squares method to Create Linear Regressions
* Assess R-Squared for all types of models
* Assess the Adjusted R-Squared for all types of models
* Create a Simple Linear Regression (SLR)
* Create a Multiple Linear Regression (MLR)
* Create Dummy Variables
* Interpret coefficients of an MLR
* Read statistical software output for created models
* Use Backward Elimination, Forward Selection, and Bidirectional Elimination methods to create statistical models
* Create a Logistic Regression
* Intuitively understand a Logistic Regression
* Operate with False Positives and False Negatives and know the difference
* Read a Confusion Matrix
* Create a Robust Geodemographic Segmentation Model
* Transform independent variables for modelling purposes
* Derive new independent variables for modelling purposes
* Check for multicollinearity using VIF and the correlation matrix
* Understand the intuition of multicollinearity
* Apply the Cumulative Accuracy Profile (CAP) to assess models
* Build the CAP curve in Excel
* Use Training and Test data to build robust models
* Derive insights from the CAP curve
* Understand the Odds Ratio
* Derive business insights from the coefficients of a logistic regression
* Understand what model deterioration actually looks like
* Apply three levels of model maintenance to prevent model deterioration
* Install and navigate SQL Server
* Install and navigate Microsoft Visual Studio Shell
* Clean data and look for anomalies
* Use SQL Server Integration Services (SSIS) to upload data into a database
* Create Conditional Splits in SSIS
* Deal with Text Qualifier errors in RAW data
* Create Scripts in SQL
* Apply SQL to Data Science projects
* Create stored procedures in SQL
* Present Data Science projects to stakeholders

Course Pathways

**Section 2: Data Science**

**Section 3: Part 1: Visualisation**

**Section 4: Introduction to Tableau**

Installing Tableau Desktop: [www.tableau.com](http://www.tableau.com)

Sample Data Source: [www.superdatascience.com/pages/training](http://www.superdatascience.com/pages/training)

Connecting Tableau to Data File: Open Tableau, on left click Connect 🡪 Text File, Select officeSupplies.csv

After opening the text file: Sorted fields are in the arranged in order of String, Date, Number for all data’s.

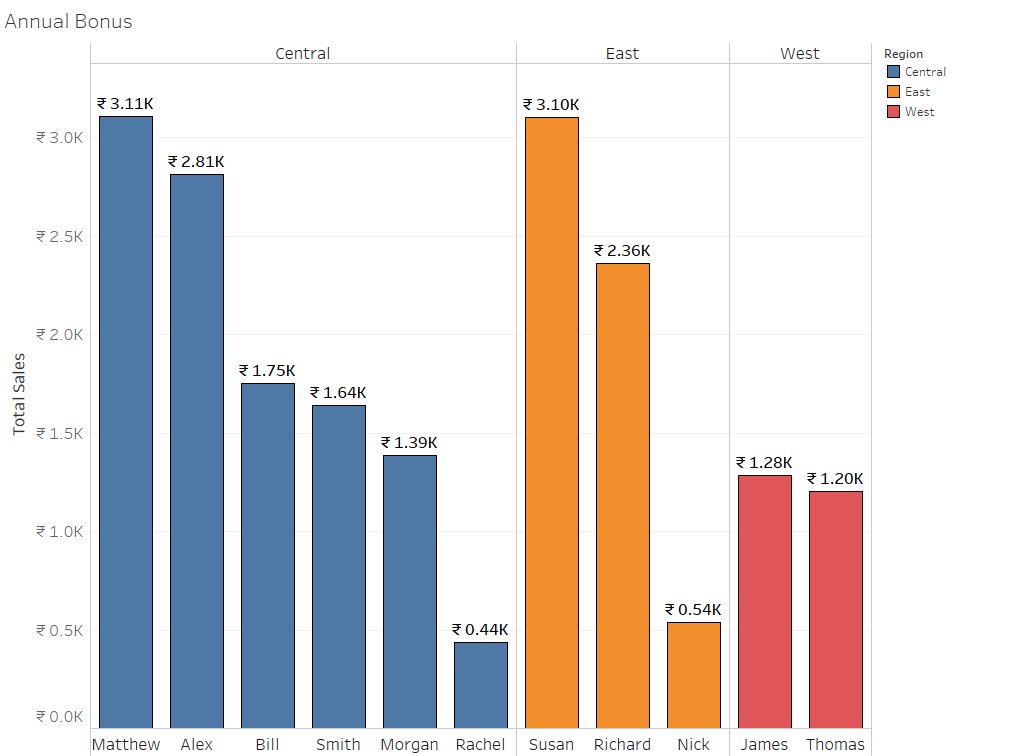
14. Creating a calculated field: Right Click on Measures, Create Calculated Field, Name the Field: Use [] to get the values of Measures. Example: [Units]\*[UnitPrice] 🡪 Click OK to get the TotalSales

15. Colors: i). Drag and Drop Dimension field on Color, this will give each value fo field a different color OR ii). Hold the Click, and pull Field, bring it on Top of the datas and press control, holding the click, a plus sign will appear and move and drop it on top of Color. (if we have multiple columns and rows, easy way)

We can edit color and use a palette

16. Label: Drag and Drop, press control and drop on to Label. If Label is not Shown, Right click on Label 🡪 Mark Label 🡪 Always Show

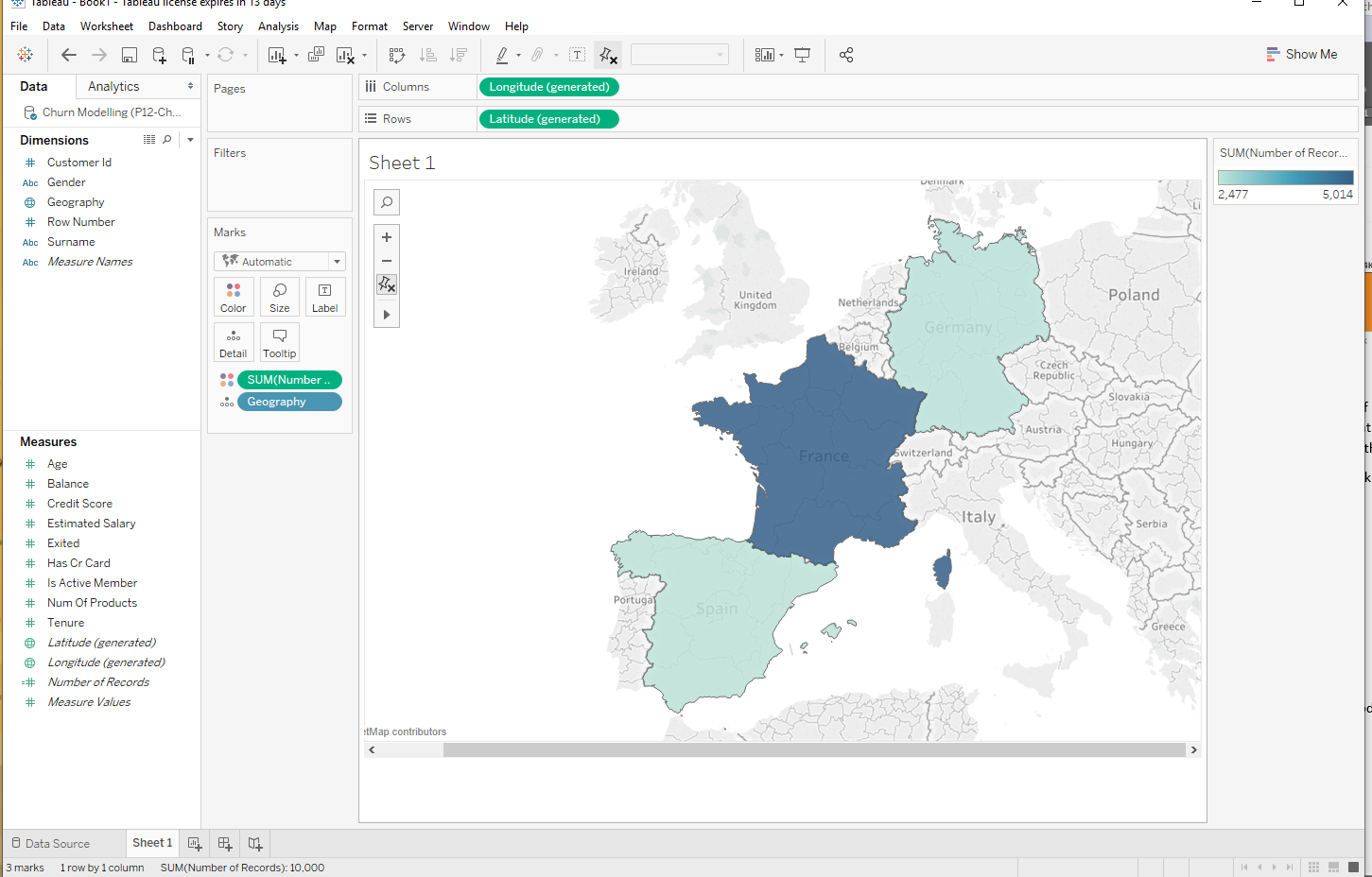
Right Click on Label or Axis, Click Format to change to currency, numbers, change decimal places

Sample: 

**Section 5: How to use Tableau for Data Mining**

Changing fields of type Strings to Location (geography), if it is an area or a location. Right click on the field, Change type to Geographic Role 🡪 Country/Region 🡪 Map will be loaded accordingly and Latitude/Longitude will be generated in the Measures column.

**TIP:** To get the most no of Values, Drag and Drop values to (like No. of records) to Color and Countries will get coloured.



* + - How to visualise an ad-hoc A-B testing in Tableau (referring Example – Chrun Modelling and Tableau)

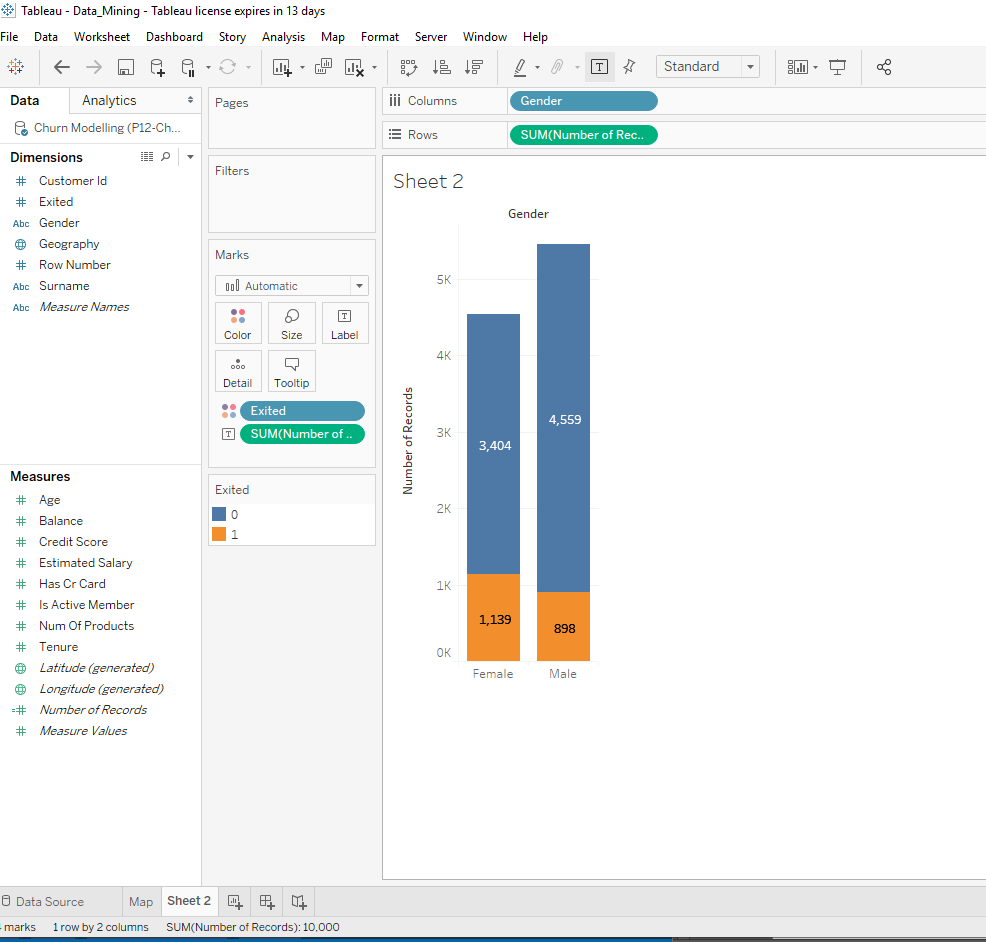
*Statistical Significance: In statistical hypothesis testing, a result has statistical significance when it is very unlikely to have occurred given the null hypothesis. More precisely, a study's defined significance level, α, is the probability of the study rejecting the null hypothesis, given that it were true; and the p-value of a result, p, is the probability of obtaining a result at least as extreme, given that the null hypothesis were true. The result is statistically significant, by the standards of the study, when p.*

* *Make informed decisions*
* *Prioritize your testing values*
* *Make more clear and Informative of all data*
* *Confirm a new design is going in the right direction*
* *Decide which version of different approaches to implement*
* *Figure out what is working best among specific UI or copy elements*
* *Learn how small changes can influence user behavior*
* *Constantly iterate a design*
* *Improve user experience over time*
* *Optimize conversion rates*

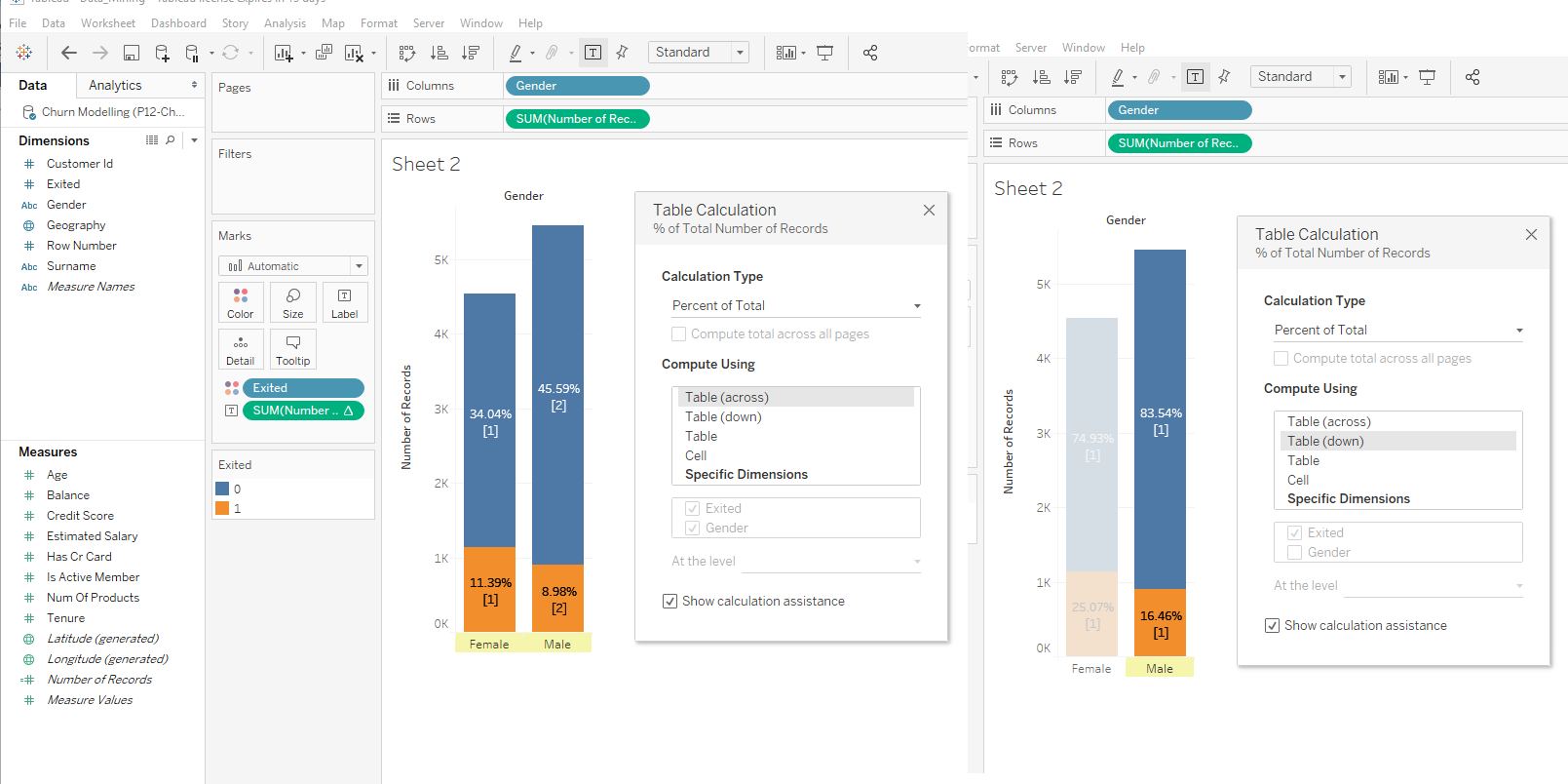
🡪 Moving from Measures to Dimensions – Exited – Did the person leave or not? By default, it was placed by Tableau in Measures, we will move it to dimensions.

A-B Test:

* 1. Drag and drop Gender to Columns and Exited to Color. This will give us 2 colors for male and female.
  2. Drop No of records into Rows: (fig: below). Drag no of records to Label to get a numbering

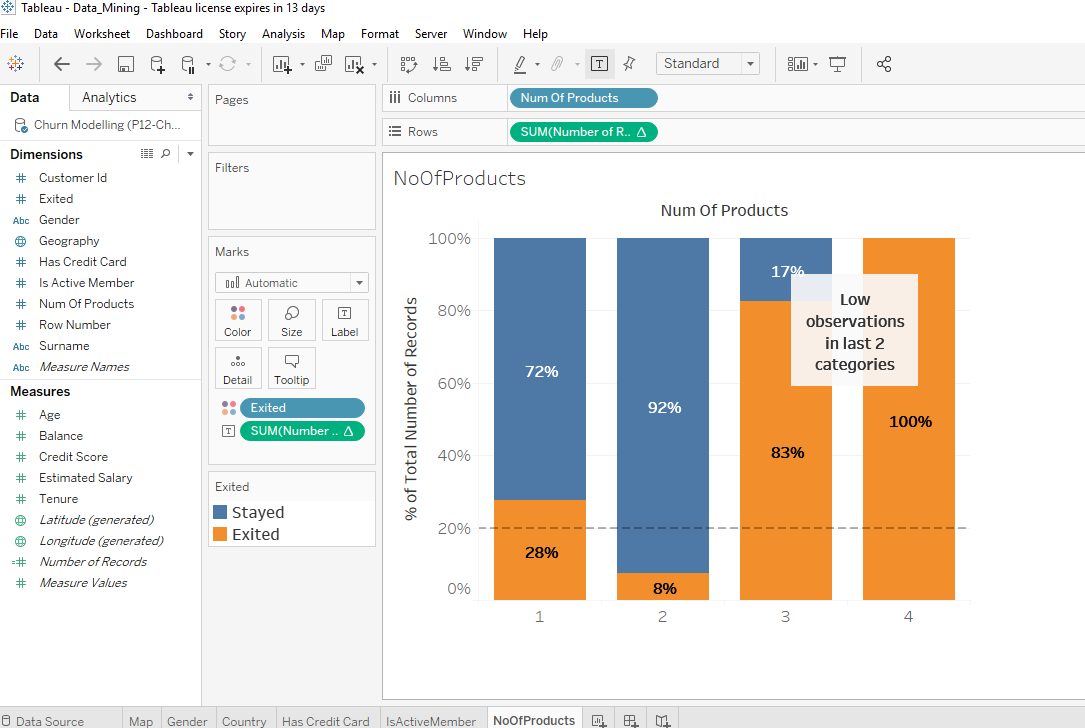


Convert absolute values to percentage: Right Click on SUM(No of records) 🡪 Add table calculation (Triangle on right represents table calculations) 🡪 Select Table(down) {Table down gives percent of a full column, regardless of n no of records. If we have 4 records, percent will be calculated according to all 4 part in a single column.} Table (across) - {will give the values of records of each part of a column, across the rows.} (fig: below)



Drag and Drop SUM(No of records) to rows, to change the axis and get a better comparing edge for records.

* Work with Aliases: Column or Row where values are 0 or 1. Right click on that column and select aliases, Change to a convenient name and press OK.
* Add a reference line: (Why - To compare results with some benchmarks). Right click on Axis and select Add reference line. Select from a list of reference lines. Select, line then entire table and provide a line value (benchmark value) and select Constant.
* Looking for Anomalies: Check for various fields with A-B testing, Vary in data, Check for Changes in values. Marking the data : Right Click on the area, select Annotate, Put a note shown in figure below.



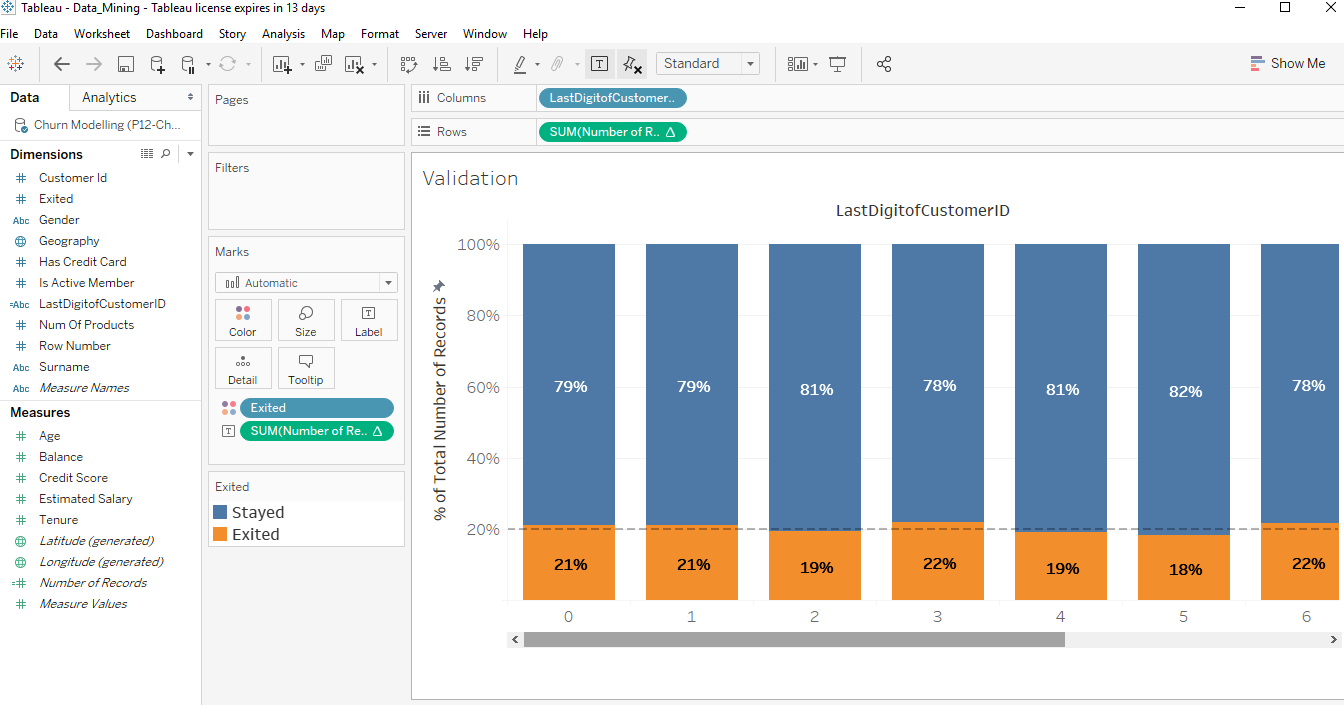
* Validate your approach/data: Find a variable for sure, that does not affect your end result. For this example: we are using Customer ID. Right Click on Customer ID 🡪 Create calculated field 🡪 Name tab as: last digit of customer id :

Expression : RIGHT(STR([Customer Id]),1). Each expression has a formula of its own.

Replace the generated formula over SUM(No. of products) to get the below detailed view.

Check for various formulas. Below is a sample from Tableau.

The below image gives a detailed view of any anomalies and will be reflected over the graph.



* Recap:
  + Connect to tableau to an excel file
  + Visualise an Ad-Hoc A-B test
  + Create Aliases (convert 1 to Yes or 0 to No)
  + Add a reference line to a chart (below average or benchmark)
  + Use Ad-Hoc A-B test to find anomalies
  + Validate your data / approach

**Section 6: Advance Data Mining with Tableau**

**Section 7: Part 2: Modelling**

**Section 8: Stats Refresher**

**Section 9: Simple Linear Regression**

**Section 10: Multiple Linear Regression**

**Section 11: Logistic Regression**

**Section 12: Building a robust geodemographic segmentation model**

**Section 13: Assessing your model**

**Section 14: Drawing insights from the model**

**Section 15: Model Maintenance**

**Section 16: Part 3: Data Preparation**

**Section 17: Business Intelligence (BI) Tools**

**Section 18: ETL Phase 1: Data Wrangling before the Load**

**Section 19: ETL Phase 2: Step-by-step guide to uploading data using SSIS**

**Section 20: Handling errors during ETL (Phase 1 & 2)**

**Section 21: SQL Programming for Data Science**

**Section 22: ETL Phase 3: Data Wrangling after the load**

**Section 23: Handling errors during ETL (Phase 3)**

**Section 24: Part 4: Communication**

**Section 25: Working with people**

**Section 26: Presenting for Data Scientists**

**Section 27: Homework Solutions**

**Section 28: Bonus Lectures**

