Seneca | SDDS

Business Analytics

Introduction to Data Visualization

BAN140 - Section NBB /NCC

Mufleh Al-Shatnawi, Ph.D., P.Eng.,

Week 3

Week Topics

Previous Week

➢Basic Data Types

➢Understanding Data

➢Discrete and continuous data

Current Week

➢Data Aggregation

➢Basic Data Analysis in Tableau

Seneca

About Data Aggregation

Click here

Introduction - How much you can find

Levels of aggregation

Factoid	Series	Multiseries	Summable Multiseries	Summary Records	Individual Transactions

Limited ability to explore and pivot   More options to explore and pivot

Introduction - levels of aggregation

Level of Aggregation	Number of metrics	Description
Factoid	Maximum context	Single data point; No drill-down
Series	One metric, across an axis	Can compare rate of change
Multiseries	Several metrics, common axis	Can compare rate of change, correlation between metrics
Summable multiseries	Several metrics, common axis	Can compare rate of change, correlation between metrics; Can compare percentages to whole
Summary records	One record for each item in a series; Metrics in other series have been aggregated somehow	Items can be compared
Individual transactions	One record per instance	No aggregation or combination; Maximum drill-down

Café Shop Example

The basic table of data, by year, looks like this

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008
Total sales	19,795	23,005	31,711	40,728	50,440	60,953	74,143	93,321	120,312
Male	12,534	16,452	19,362	24,726	28,567	31,110	39,001	48,710	61,291
Female	7,261	6,553	12,349	16,002	21,873	29,843	35,142	44,611	59,021
Regular	9,929	14,021	17,364	20,035	27,854	34,201	36,472	52,012	60,362
Decaf	6,744	6,833	10,201	13,462	17,033	19,921	21,094	23,716	38,657
Mocha	3,122	2,151	4,146	7,231	5,553	6,831	16,577	17,593	21,293

These are completely made up coffee data, BTW.

Seneca

Factoid

➢ A factoid, or trivial fact, is a single piece of information that emphasizes a particular point of view, idea, or detail.

➢ A factoid does not allow for any further statistical analysis.

➢ A factoid is calculated from source data

36.7% of coffee in 2000 was consumed by women.

Series
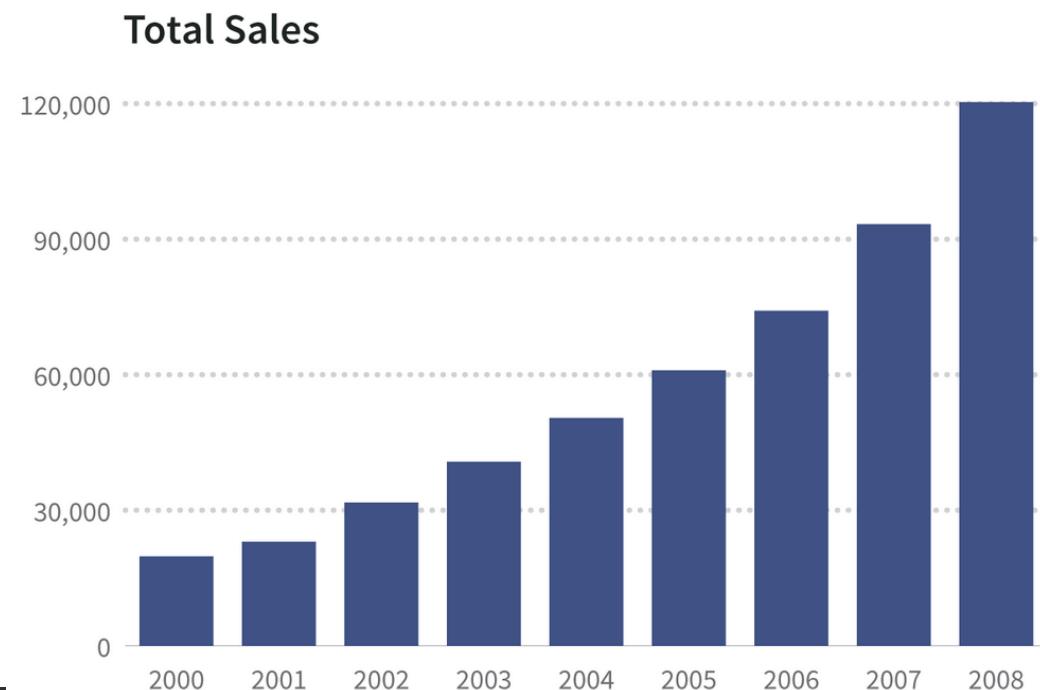
➢This is **one type** of information (the dependent variable) **compared to another** (the independent variable). Often, the independent variable is time.

Year	2000	2001	2002	2003
Total sales	19,795	23,005	31,711	40,728

More about Series

➢ A series can also be some other set of **continuous data**, such as **temperature**.

➢ A series of **non-contiguous**, **but related, information in a category**, such as major car brands, types of dog, vegetables, or the mass of planets in the solar system

Total Sales
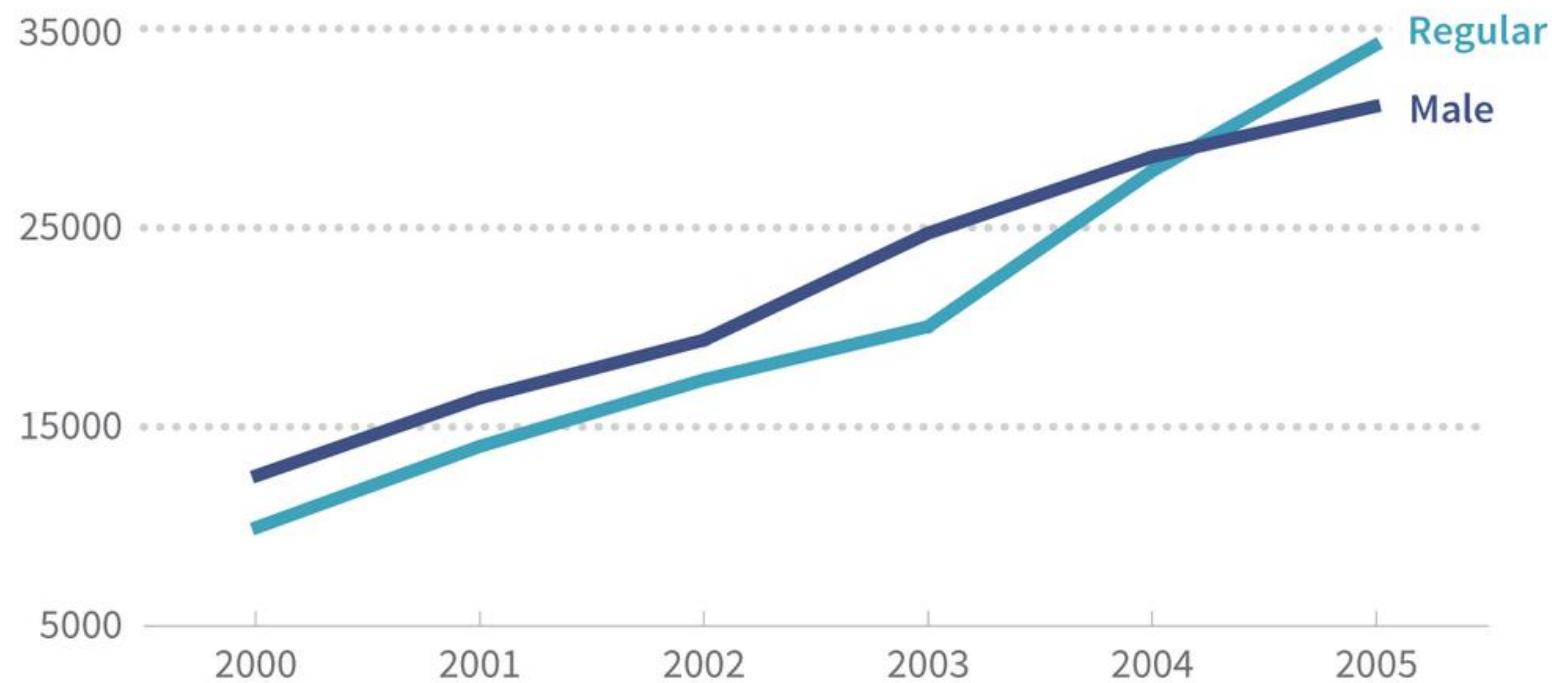
Multiseries

➢ A multiseries dataset has several pieces of **dependent information** and one piece of independent information.

Year	2000	2001	2002	2003	2004	2005
Male	12,534	16,452	19,362	24,726	28,567	31,110
Regular	9,929	14,021	17,364	20,035	27,854	34,201

It Good to draw but you can not conclude
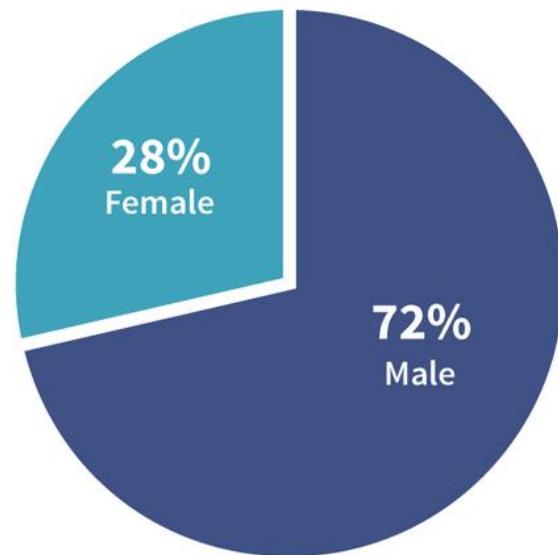
➢A Multiseries data are simply several series on one chart or table. We can show them together, but we can't meaningfully stack or combine them.
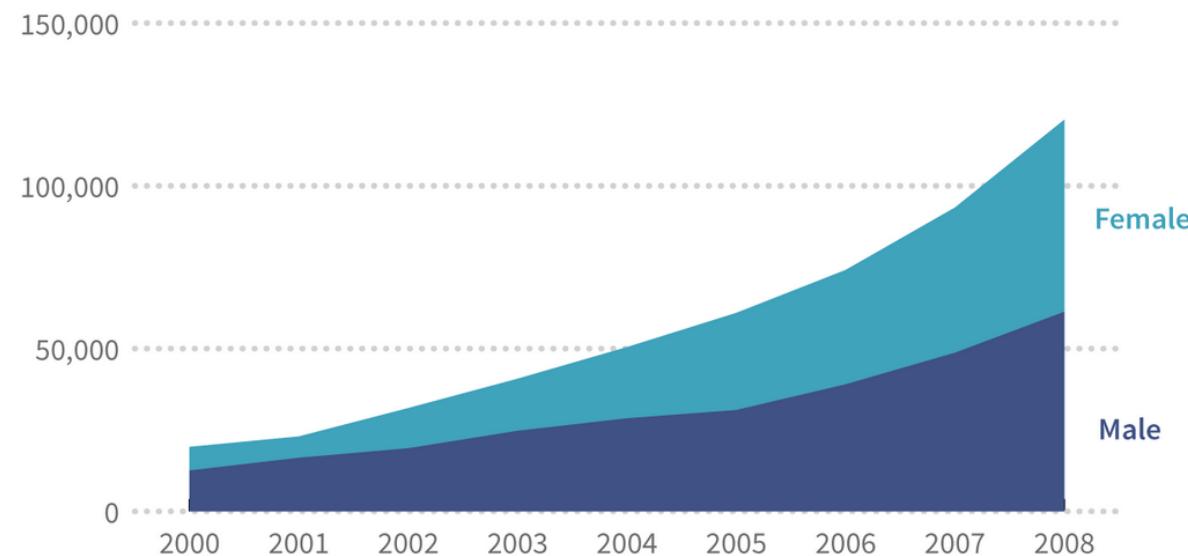
Summable Multiseries

➢A summable multiseries is a particular statistic (gender, type of coffee) segmented into subgroups.

➢Given information about coffee drinker, we can add these together to make broader observations about total consumption. For one thing, we can display percentages.

Coffee consumption by gender in 2001

28% Female

72% Male

Total cups of coffee, by gender

Challenge with Summable Multiseries

Year	2000	2001	2002	2003	2004
Male	12534	16452	19362	24726	28567
Female	7261	6553	12349	16002	21873
Regular	9929	14021	17364	20035	27854
Decaf	6744	6833	10201	13462	17033
Mocha	3122	2151	4146	7231	5553

Seneca

Wrong Conclusion/inference

➢This is a common (and important) mistake. Many people are tempted to say:

   ➢**36.7% of cups sold in 2000 were sold to women.**

   ➢And there were 9,929 cups of regular sold in 2000.

   ➢Therefore, 3,642.5 cups of regular were sold to women.

➢But this is wrong. This type of inference can only be made when you know that one category (coffee type) is **evenly distributed across another (gender)**.

Summary Records

➢Summary records looks like the kind of data a point-of-sale system at a café might generate. It includes a column of categorical information (gender, where there are two possible types) and subtotals for each type of coffee. It also includes the totals by the cup for those types.

Name	Gender	Regular	Decaf	Mocha	Total
Bob Smith	M	2	3	1	6
Jane Doe	F	4	0	0	4
Dale Cooper	M	1	2	4	7
Mary Brewer	F	3	1	0	4
Betty Kona	F	1	0	0	1
John Java	M	2	1	3	6
Bill Bean	M	3	1	0	4
Jake Beatnik	M	0	0	1	1
Totals	5M, 3F	16	8	9	33
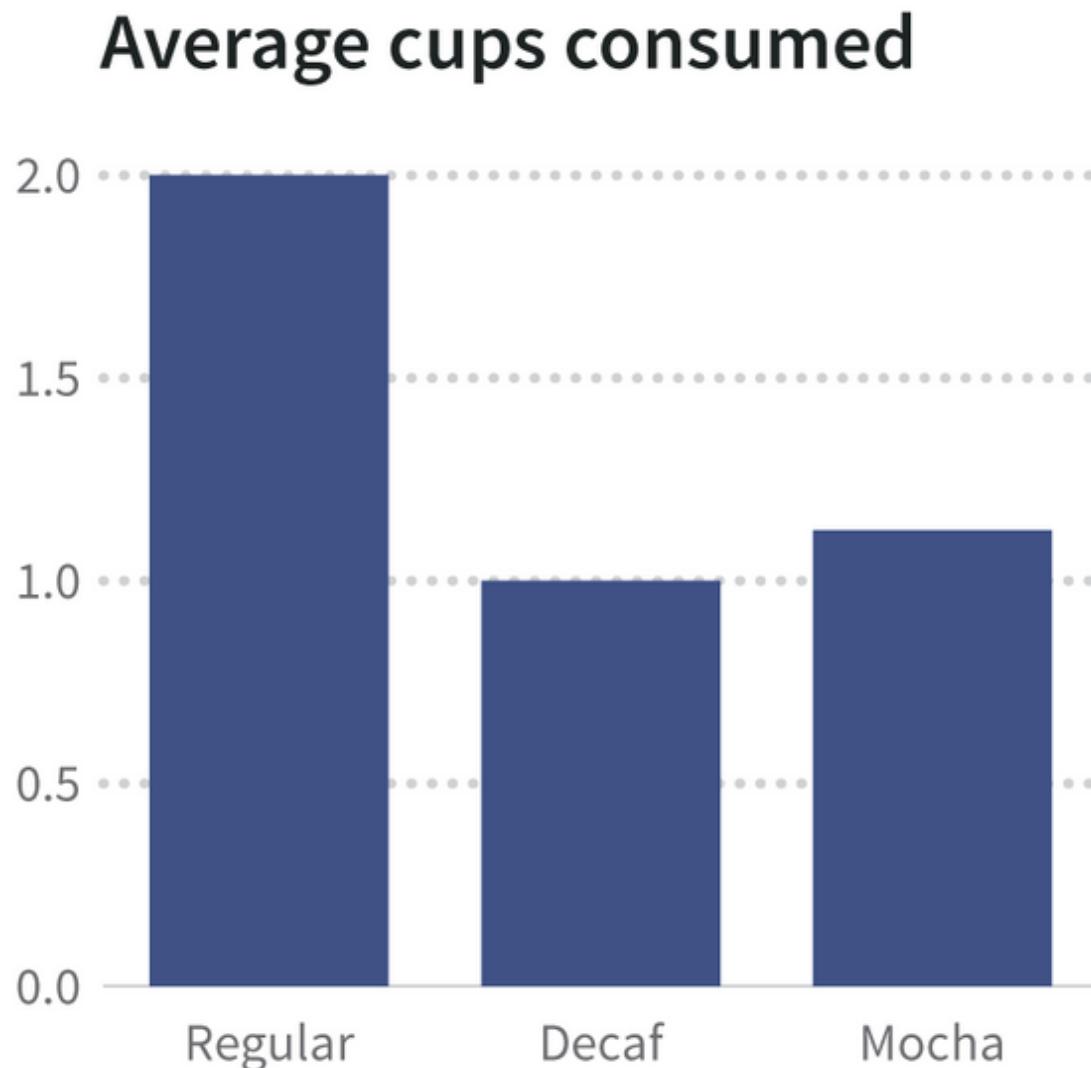
We can ask question and try to answer it
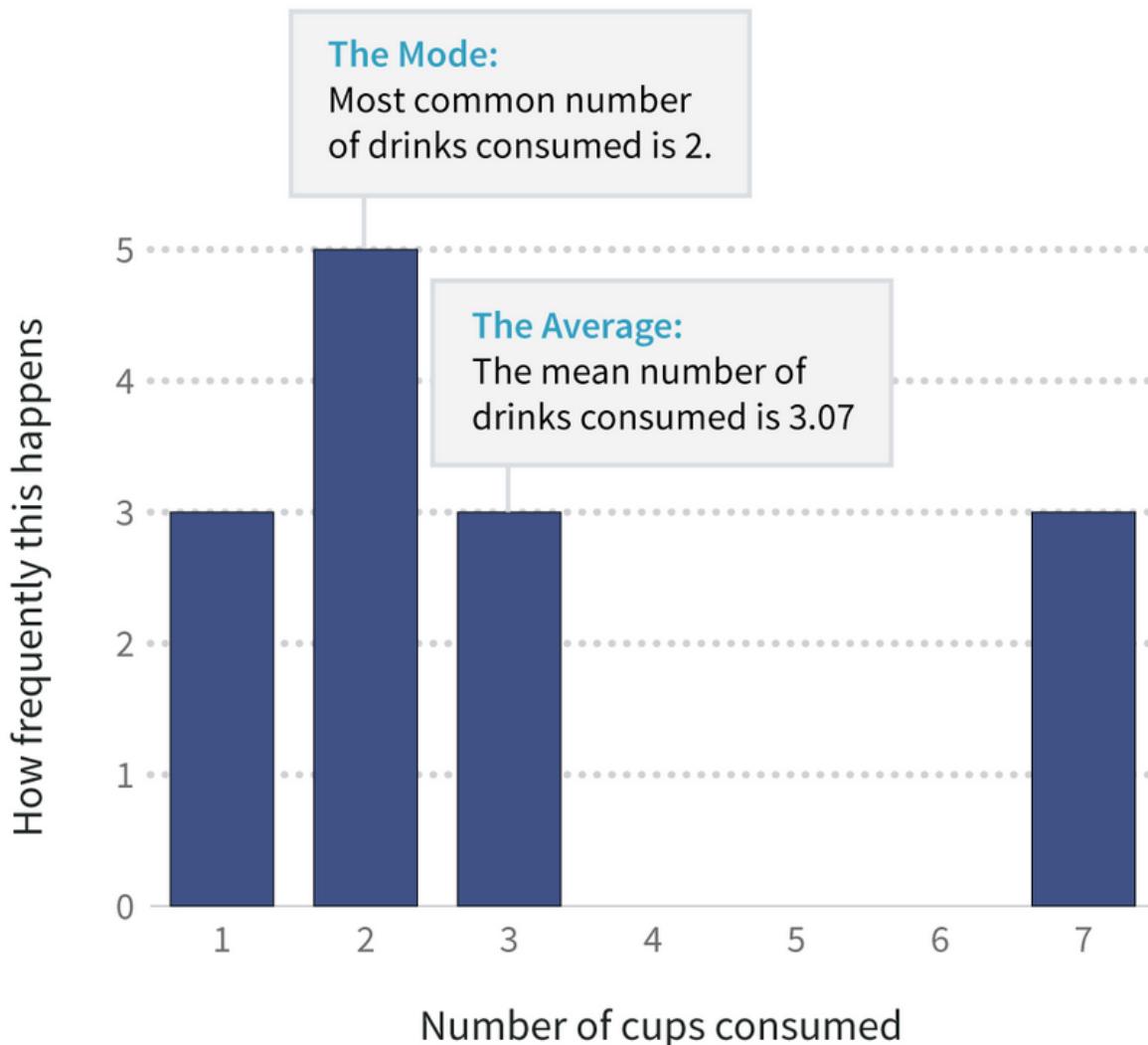
Row Labels	Average of Regular	Average of Decaf	Average of Mocha
F	2.67	0.33	0.00
M	2.00	1.75	2.00
Grand Total	2.29	1.14	1.14

Using Visualization to Reveal Underlying Variance

Name	Regular	Decaf	Mocha
Bob Smith	2	3	1
Jane Doe	4	0	0
Dale Cooper	1	2	4
Mary Brewer	3	1	0
Betty Kona	1	0	0
John Java	2	1	3
Bill Bean	3	1	0
Jake Beatnik	0	0	1
Totals	16	8	9
Averages	2	1	1.125

Average cups consumed

A better Visualization to Reveal Underlying Variance

When you have raw data, you can see the exceptions and outliers, and tell a more accurate story.

Individual Transactions

➢Transactional records capture things about a specific event.

➢There's no aggregation of the data along any dimension like someone's name (though their name may be captured).

➢It's not rolled up over time; it's instantaneous.

These transactions can be aggregated by any column. They can be cross-referenced by those columns. The timestamps can also be aggregated into buckets (hourly, daily, or annually). Ultimately, the initial dataset we saw of coffee consumption per year results from these raw data.

Timestamp	Name	Gender	Coffee
17:00	Bob Smith	M	Regular
17:01	Jane Doe	F	Regular
17:02	Dale Cooper	M	Mocha
17:03	Mary Brewer	F	Decaf
17:04	Betty Kona	F	Regular

TABLEAU DESKTOP

➢ Desktop is the design tool for creating visual analytics and dashboards.

➢ Connect to different data source

➢ Create a data visualization

➢ Combine multiple visualizations into an interactive dashboard ("dashboard")

Tableau Software

➢ [Wikipedia] Tableau Software, Inc.
  ➢ An American *interactive* data visualization software company
  ➢ Founded on January 2003 by Christian Chabot, Pat Hanrahan and Chris Stolte (researchers at Stanford University)
  ➢ Focused on business intelligence.
  ➢ On August 1, 2019, Salesforce.com acquired Tableau.

➢ Products
  ➢ Tableau Desktop
  ➢ Tableau Server
  ➢ Tableau Online
  ➢ Tableau Public
  ➢ Tableau Prep Builder, Tableau Reader

➢ Ash Patel @Seneca: View Video Tutorial (16:46 min)

Tableau Software

Data visualization company

tableau.com

Tableau Software is an interactive data visualization software company founded in January 2003 by Christian Chabot, Pat Hanrahan and Chris Stolte, in Mountain View, California. The company is currently headquartered in Seattle, Washington, United States focused on business intelligence. Wikipedia

Founded: January 2003, Mountain View, California, United States

Parent organization: Salesforce

CEO: Adam Selipsky (Sep. 16, 2016–)

Headquarters: Seattle, Washington, United States

Revenue: 1.2 billion USD (2018)

Founders: Pat Hanrahan, Christian Chabot, Andrew Beers, Chris Stolte

Tableau File Type

You can save and share data using a variety of different file types.

FILE TYPE (FILE EXTENSION)	SIZE	USE CASE	INCLUDES
Tableau Workbook (twb)	Small	Tableau's default way to save work.	Information to visualize data. No source data.
Tableau Datasource (tds)	Small	Accessing frequently-used datasources.	Server address, pass-word, and other meta-data related to the datasource.
Tableau Bookmark (tbm)	Normally small	Sharing work-sheets from one workbook to another.	Information to visualize and the datasource if the source workbook is a packaged workbook.
Tableau Data Extract (tde)	Potentially large	Improves perfor-mance. Enables more functions.	Source data as filtered and aggregated during extract.
Tableau Packaged Workbook (twbx)	Potentially large	Sharing with Tableau Reader or those without access to the source data.	Extracted data and workbook information to build visualizations.

Link Visualization File to Source Data

Tutorial: Get Started with Tableau Desktop

Data Aggregation in Tableau

Seneca

References and Resources

➢[Knaflic] Cole Nussbaumer Knaflic, **Storytelling with Data: A Data Visualization Guide for Business Professionals**, Wiley, 2017
  - ➢ Available online through Seneca Libraries: https://senecacollege-primo.hosted.exlibrisgroup.com/permalink/f/t3376v/01SENC_ALMA5146374280003226

➢ [Ryan] Lindy Ryan, **Visual Data Storytelling with Tableau**, Pearson Addison-Wesley, 2018
  - ➢ Available online through Seneca Libraries: https://senecacollege-primo.hosted.exlibrisgroup.com/permalink/f/t3376v/01SENC_ALMA5167006190003226

➢ [Healy] Kieran Healy, **Data Visualization: A Practical Introduction**, Princeton University Press, 2018.
  - ➢ Available (hardcopy) at Seneca Libraries: https://senecacollege-primo.hosted.exlibrisgroup.com/permalink/f/t3376v/01SENC_ALMA2172469250003226

➢**A Reader on Data Visualization:** https://mschermann.github.io/data_viz_reader/

➢**Data visualization:** https://en.wikipedia.org/wiki/Data_visualization

➢**Section 5: Data concepts** https://www.statcan.gc.ca/eng/dli/guide/section5