ECE5984 – Applications of Machine Learning
Lecture 4 – Data and Data Exploration

Creed Jones, PhD

BRADLEY DEPARTMENT
OF ELECTRICAL & COMPUTER ENGINEERING

VIRGINIA TECH.

Course Updates

- Quiz 1 is today
  - Noon Thursday to 3 AM Friday, EST (long period this time)
  - 20 minute time limit
- Next quiz on February 10
  - Covers lectures 4-7
- At the end of the semester, I will replace your lowest quiz grade with your next lowest grade
- HW1
  - Due on Feb 8
  - Submit via Canvas

Today’s Objectives

Chapter 2 – Data to Insights

- 2.1 Converting Business Problems into Analytics Solutions
- 2.2 Assessing Feasibility
- 2.3 An Analytics Base Table
- 2.4 Features

Descriptive Statistics on a Dataset

Tableau

CHAPTER 2 – DATA TO INSIGHTS

Using data to generate insights or provide answers requires that we clearly understand the problem

- Fact: we only get paid to do machine learning because we help the organization we are part of
- Fact: all organizations can benefit from ML
- Fact: most organizations are new to using ML and aren't always skilled at thinking in terms of how to use it
- Fact: it's often up to us to understand the real issues to be addressed and come up with creative ways to solve the problems

- Conclusion: we as ML practitioners have to understand the *business problem* and define a technical solution to it

Converting a business problem into an analytics solution involves answering the following key questions:

1. What is the business problem?
2. What are the goals that the business wants to achieve?
3. How does the business currently work?
4. In what ways could a predictive analytics model help to address the business problem?

Case Study: Motor Insurance Fraud

In spite of having a fraud investigation team that investigates up to 30% of all claims made, a motor insurance company is still losing too much money due to fraudulent claims.

- What predictive analytics solutions could be proposed to help address this business problem?

Converting a business problem into an analytics solution involves answering the following key questions:

1. What is the business problem?
2. What are the goals that the business wants to achieve?
3. How does the business currently work?
4. In what ways could a predictive analytics model help to address the business problem?

Case Study: Motor Insurance Fraud

In spite of having a fraud investigation team that investigates up to 30% of all claims made, a motor insurance company is still losing too much money due to fraudulent claims.

- What predictive analytics solutions could be proposed to help address this business problem?

- Potential analytics solutions include:
  - Claim prediction
  - Member prediction
  - Application prediction
  - Payment prediction

Question	We want to help students that will struggle in a given course	Targeted marketing – pushing ads out to likely customers	Detect fruit that has hidden spoiled patches inside
What is the business problem?			
What are the goals that the business wants to achieve?			
How does the business currently work?			
In what ways could a predictive analytics model help to address the business problem?			

Evaluating the feasibility of a proposed analytics solution involves considering the following questions:

1. Is the data required by the solution available, or could it be made available?

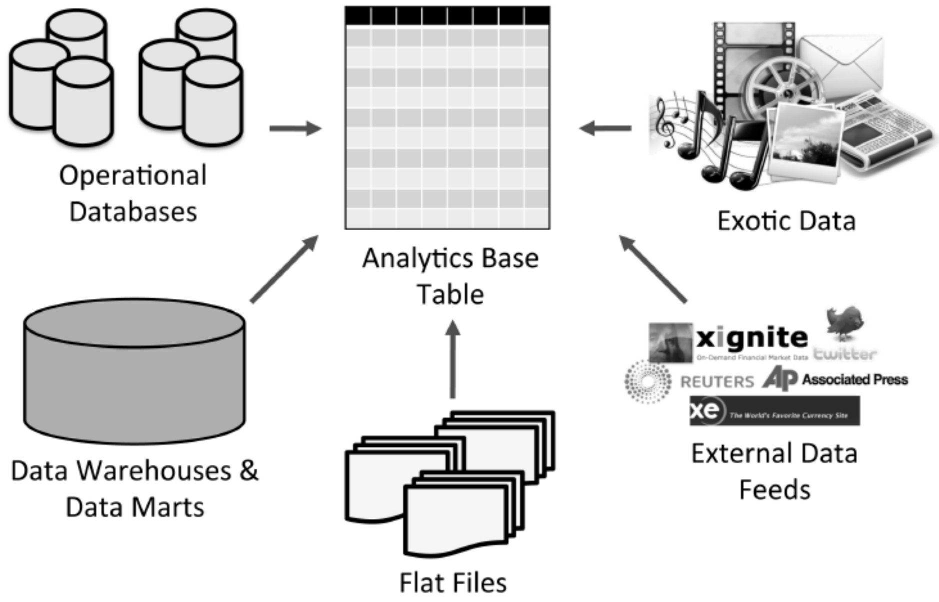
2. What is the capacity of the business to utilize the insights that the analytics solution will provide?

Evaluating the feasibility of a proposed analytics solution involves considering the following questions:

1. Is the data required by the solution available, or could it be made available?
   - Constraints may be technical, temporal, legal or economic
   - What if I have *most* of the data I need for *most* instances?
   - Sometimes proxy variables can provide a suboptimal but sufficient solution

2. What is the capacity of the business to utilize the insights that the analytics solution will provide?
   - Again, constraints may be technical, temporal, legal or economic
   - May also be related to culture or business model

In the Analytic dataset or *Analytics Base Table*, each row is an instance or example and each column is an ID, descriptive feature or target variable



- IDs are used to distinguish instances, subjects or other data not used for modeling
- Target variables (one or several) are the outputs or results that we want the model to estimate or predict
- Descriptive features are suitable for modeling
  - May be of different types
  - May have missing or invalid values
  - Range of the data may be an issue
  - May be calculated

Many data types can be used in ML systems – when discussing specific modeling techniques, we will need to see what feature types are supported

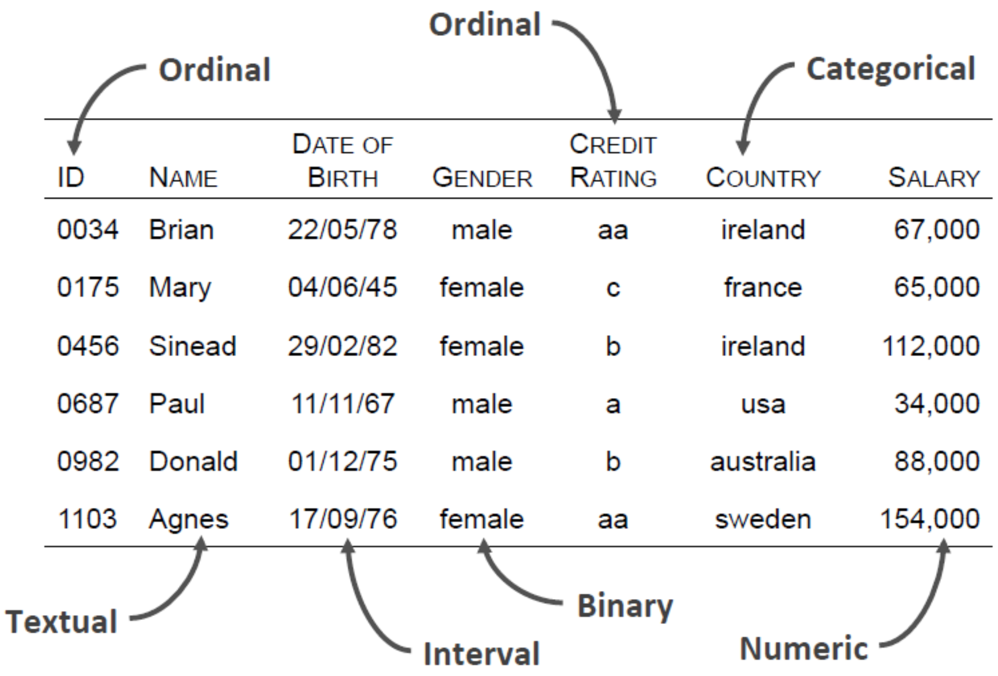
- **Numeric:** True numeric values that allow arithmetic operations (e.g., price, age)
- **Interval:** Values that allow ordering and subtraction, but do not allow other arithmetic operations (e.g., date, time)
- **Ordinal:** Values that allow ordering but do not permit arithmetic (e.g., size measured as small, medium, or large)
- **Categorical:** A finite set of values that cannot be ordered and allow no arithmetic (e.g., country, product type)
- **Binary:** A set of just two values (e.g., present/absent)
- **Textual:** Free-form, usually short, text data (e.g., name, address)

Look at an example of some types of features in a small analytic data set

Employee ID	Salary	Hire Date	Job Level	Department	Work from Home	Manager	Last Name	Expects Raise?
Numeric	*Numeric*	*Interval*	*Ordinal*	*Categorical*	*Binary*	*Textual*	*Textual*	*Binary*
ID	*Feature*	*Feature*	*Feature*	*Feature*	*Feature*	*Feature*	*ID*	*Target*
1002353	$ 88,300	1-Jan-18	5	Sales	No	Smith	Tinker	No
1013424	$ 91,500	16-Jun-12	5	Sales	Yes	Allen	Evers	No
1006777	$ 82,000	1-Sep-17	4	Accounting	No	Rao	Chance	Yes
1000835	$ 111,300	3-Jan-13	6	R&D	No	Baker	Casey	Yes

- Numeric: True numeric values that allow arithmetic operations (e.g., price, age)
- Interval: Values that allow ordering and subtraction, but do not allow other arithmetic operations (e.g., date, time)
- Ordinal: Values that allow ordering but do not permit arithmetic (e.g., size measured as small, medium, or large)
- Categorical: A finite set of values that cannot be ordered and allow no arithmetic (e.g., country, product type)
- Binary: A set of just two values (e.g., present/absent)
- Textual: Free-form, usually short, text data (e.g., name, address)

When selecting features, we must consider:
- Data availability
- Timing
- Type
- Longevity



ID	Name	Date of Birth	Gender	Credit Rating	Country	Salary
0034	Brian	22/05/78	male	aa	ireland	67,000
0175	Mary	04/06/45	female	c	france	65,000
0456	Sinead	29/02/82	female	b	ireland	112,000
0687	Paul	11/11/67	male	a	usa	34,000
0982	Donald	01/12/75	male	b	australia	88,000
1103	Agnes	17/09/76	female	aa	sweden	154,000

- It's common to wish we had access to some feature that is not *available*
- Data must be available to the model in time to be used
- Some data elements become obsolete
  - People move
  - Economic changes
  - New diagnoses

We typically use a mix of *raw* and *derived* features for modeling

There are a number of common derived feature types:

- Aggregates are calculations (sum, mean, max, etc.) over a group or time period
- Flags are binary indications of presence or absence of some attribute
  - Often we convert categorical variables into a set of flags
- Ratios between features are often useful
- Mappings are conversions of numerical features (ounces) into categorical features (small, medium and large)
- Groupings collect many related categories into fewer higher-level categories
  - Group “El Salvador, Panama, Nicaragua” into “Central America”

When defining or selecting features, there are some particular sorts of quantities that will often have predictive power

For a model predicting human behavior (consumer actions, for example):

- Prediction Subject Details
- Demographics
- Financial
- Residence
- Usage
- Changes in Usage
- Special Usage
- Lifecycle Phase
- Network Link

In other problem domains, other concepts are often useful:

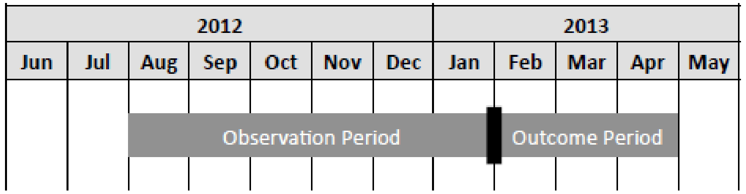
- Geographic spread
  - Disease modeling
- Global and national economic indices
  - Financial modeling
- Weather / season
- Social media activity
- News coverage
- Landmark events
  - 9/11

Many of the predictive models that we build are *propensity* models, which inherently have a temporal element

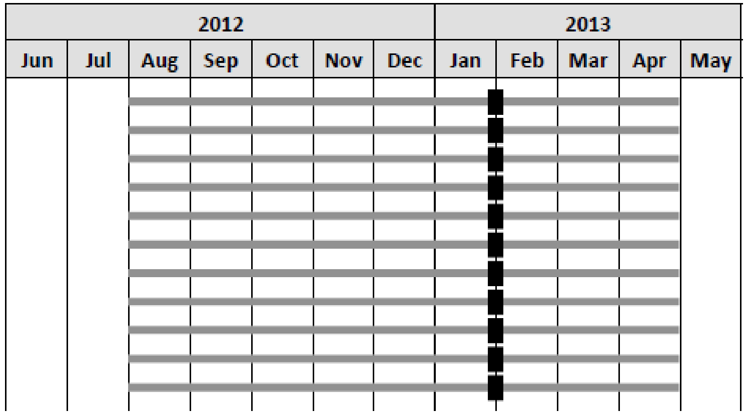
For propensity modeling, there are two key periods:

- the observation period
- the outcome period

- Sometimes the observation and outcome period are measured over the same time for all predictive subjects



(a) Observation period and outcome period

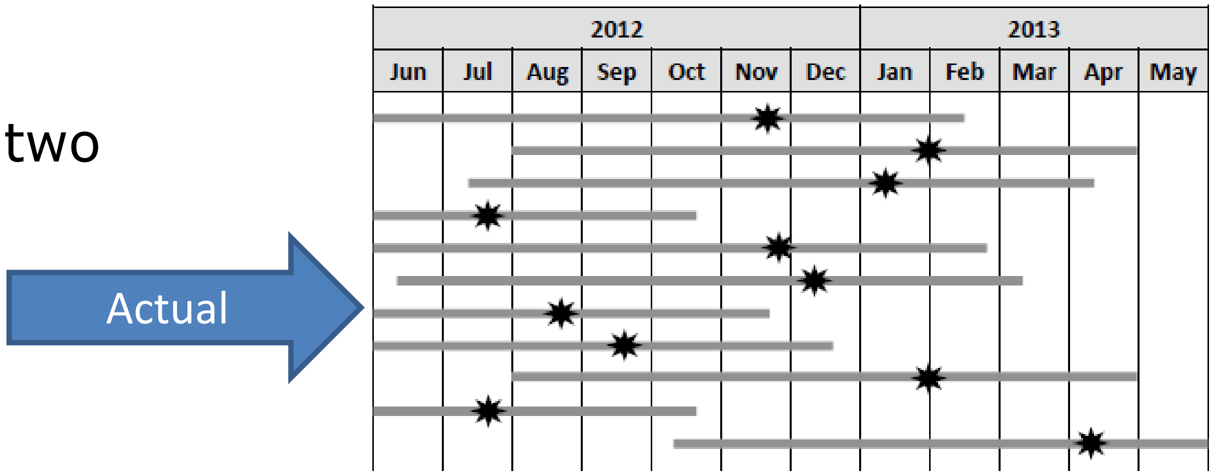


(b) Observation and outcome periods for multiple customers (each line represents a customer)

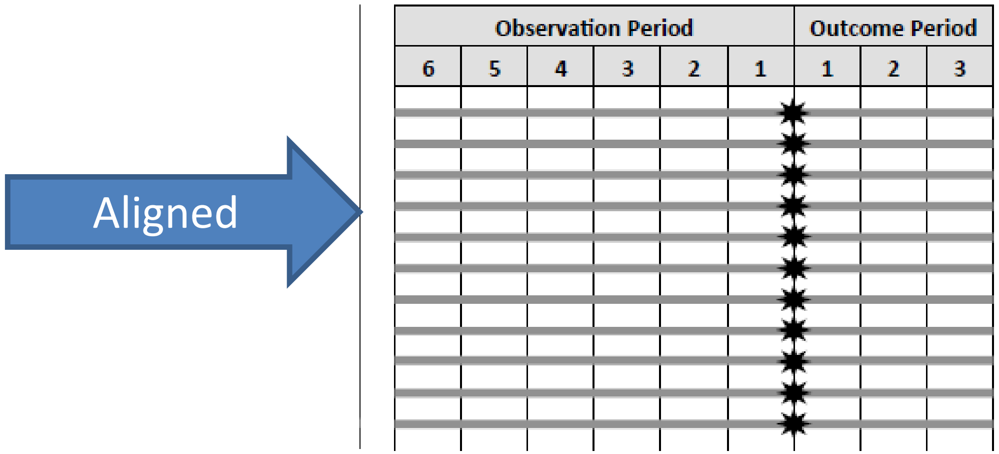
Many of the predictive models that we build are *propensity* models, which inherently have a temporal element

For propensity modeling, there are two key periods:

- the observation period
- the outcome period



- Sometimes the observation and outcome period are measured over the same time for all predictive subjects
- Often the observation period and outcome period will be measured over different dates for each prediction subject.



We often are restricted in selection of data sources or timeframes by legal constraints

There are significant differences in legislation in different jurisdictions, but a couple of key relevant principles almost always apply:

1. Anti-discrimination legislation
2. Data protection legislation
   (HIPAA, FERPA, etc.)

There are principles that we obey in our work; specific practice depends on where you are and what field you are working in –

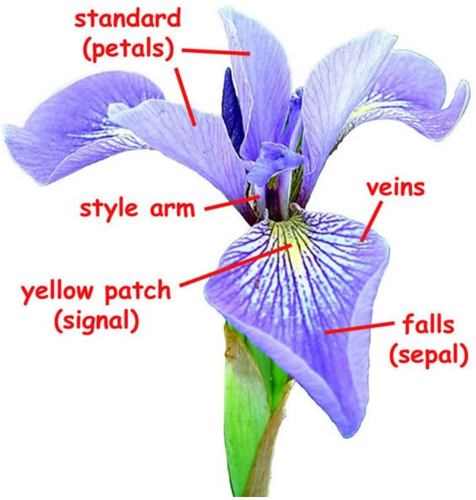
- The **collection limitation principle**
- The **purpose specification principle**
- The **use limitation principle**

DESCRIPTIVE STATISTICS FOR A DATASET

It's very useful to examine the basic descriptive statistics on an analytic table – but keep in mind that they are most descriptive of linear relationships

sepal_length	sepal_width	petal_length	petal_width	species
4.3	3	1.1	0.1	1
4.4	2.9	1.4	0.2	1
4.4	3	1.3	0.2	1
4.4	3.2	1.3	0.2	1
4.5	2.3	1.3	0.3	1
4.6	3.1	1.5	0.2	1
4.6	3.2	1.4	0.2	1
4.6	3.4	1.4	0.3	1
4.6	3.6	1	0.2	1
4.7	3.2	1.3	0.2	1

- The "iris" dataset is a classic in pattern recognition
- Three types of iris flowers
- 150 individual samples with four measures



- Used to explore methods for identifying the species from measurements

Examine descriptive statistics on the iris dataset

Statistics	sepal_length	sepal_width	petal_length	petal_width
Mean	5.843333333	3.054	3.758666667	1.198666667
Min	4.3	2	1	0.1
Max	7.9	4.4	6.9	2.5
Range	3.6	2.4	5.9	2.4
Median	5.8	3	4.35	1.3
Mode	5	3	1.5	0.2
Variance	0.685693512	0.188004027	3.113179418	0.582414318
Std Deviation	0.828066128	0.433594311	1.76442042	0.763160742
Quartile 1	5.1	2.8	1.575	0.3
Quartile 2	5.8	3	4.35	1.3
Quartile 3	6.4	3.3	5.1	1.8

COVARIANCE			
0.685693512	-0.039268456	1.273682327	0.516903803
-0.039268456	0.188004027	-0.321712752	-0.117981208
1.273682327	-0.321712752	3.113179418	1.296387472
0.516903803	-0.117981208	1.296387472	0.582414318

Examine descriptive statistics on the iris dataset

Statistics	sepal_length	sepal_width	petal_length	petal_width
Mean	5.843333333	3.054	3.758666667	1.198666667
Min	4.3	2	1	0.1
Max	7.9	4.4	6.9	2.5
Range	3.6	2.4	5.9	2.4
Median	5.8	3	4.35	1.3
Mode	5	3	1.5	0.2
Variance	0.685693512	0.188004027	3.113179418	0.582414318
Std Deviation	0.828066128	0.433594311	1.76442042	0.763160742
Quartile 1	5.1	2.8	1.575	0.3
Quartile 2	5.8	3	4.35	1.3
Quartile 3	6.4	3.3	5.1	1.8

COVARIANCE			
0.685693512	-0.039268456	1.273682327	0.516903803
-0.039268456	0.188004027	-0.321712752	-0.117981208
1.273682327	-0.321712752	3.113179418	1.296387472
0.516903803	-0.117981208	1.296387472	0.582414318

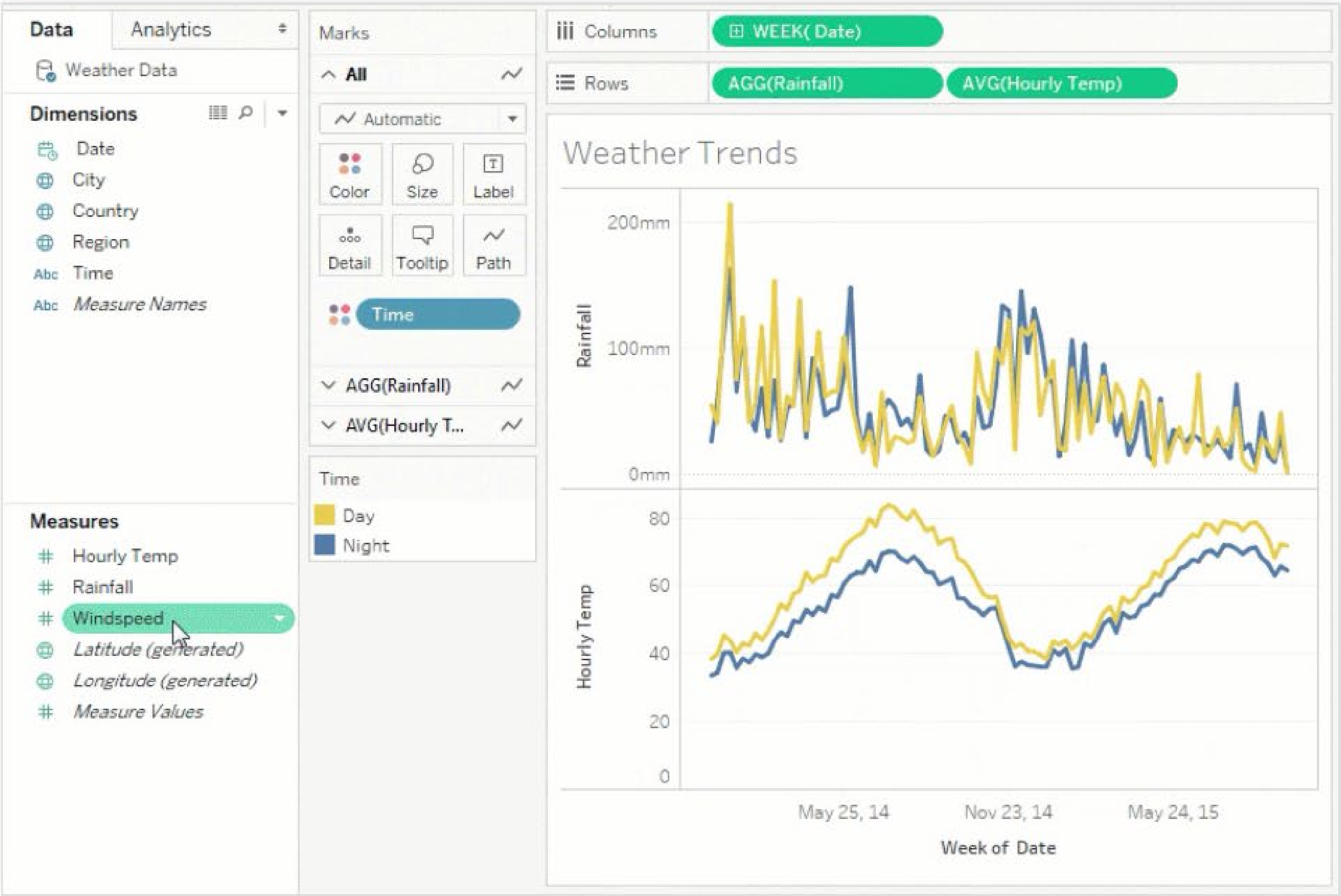
TABLEAU

Tableau is a common tool for data exploration and visualization

Get actionable insights fast

Leave chart builders behind. Live visual analytics fuel unlimited data exploration. Interactive dashboards help you uncover hidden insights on the fly. Tableau harnesses people's natural ability to spot visual patterns quickly, revealing everyday opportunities and eureka moments alike.

START A FREE TRIAL ⟶



There are some usual steps in performing *exploratory data analysis* (EDA) using Tableau

1. Connect to one or more data sources
   1. Many different formats – Excel, DB, flat text file, cloud…
   2. More than one table can be joined
2. Create a worksheet
   1. Variables (columns in the dataset) are listed on the left; many attributes are imputed
3. Explore relationships using available table, graph and plot types
   1. Bar charts, scatter plots, pie charts, histograms, heat maps, geographic…
4. There are many possible functions and enhancements of variables
   1. Groupings, binnings
   2. Aggregating functions (mean, max, count…)
   3. More complex functions can be written

In Tableau, it's important to understand the difference between *measures* and *dimensions* – and *discrete* and *continuous* quantities

Data fields are made from the columns in your data source. Each field is automatically assigned a data type (such as integer, string, date), and a role: Discrete Dimension or Continuous Measure (more common), or Continuous Dimension or Discrete Measure (less common).

- *Dimensions* contain qualitative values (such as names, dates, or geographical data). You can use dimensions to categorize, segment, and reveal the details in your data. Dimensions affect the level of detail in the view.
- *Measures* contain numeric, quantitative values that you can measure. Measures can be aggregated. When you drag a measure into the view, Tableau applies an aggregation to that measure (by default).

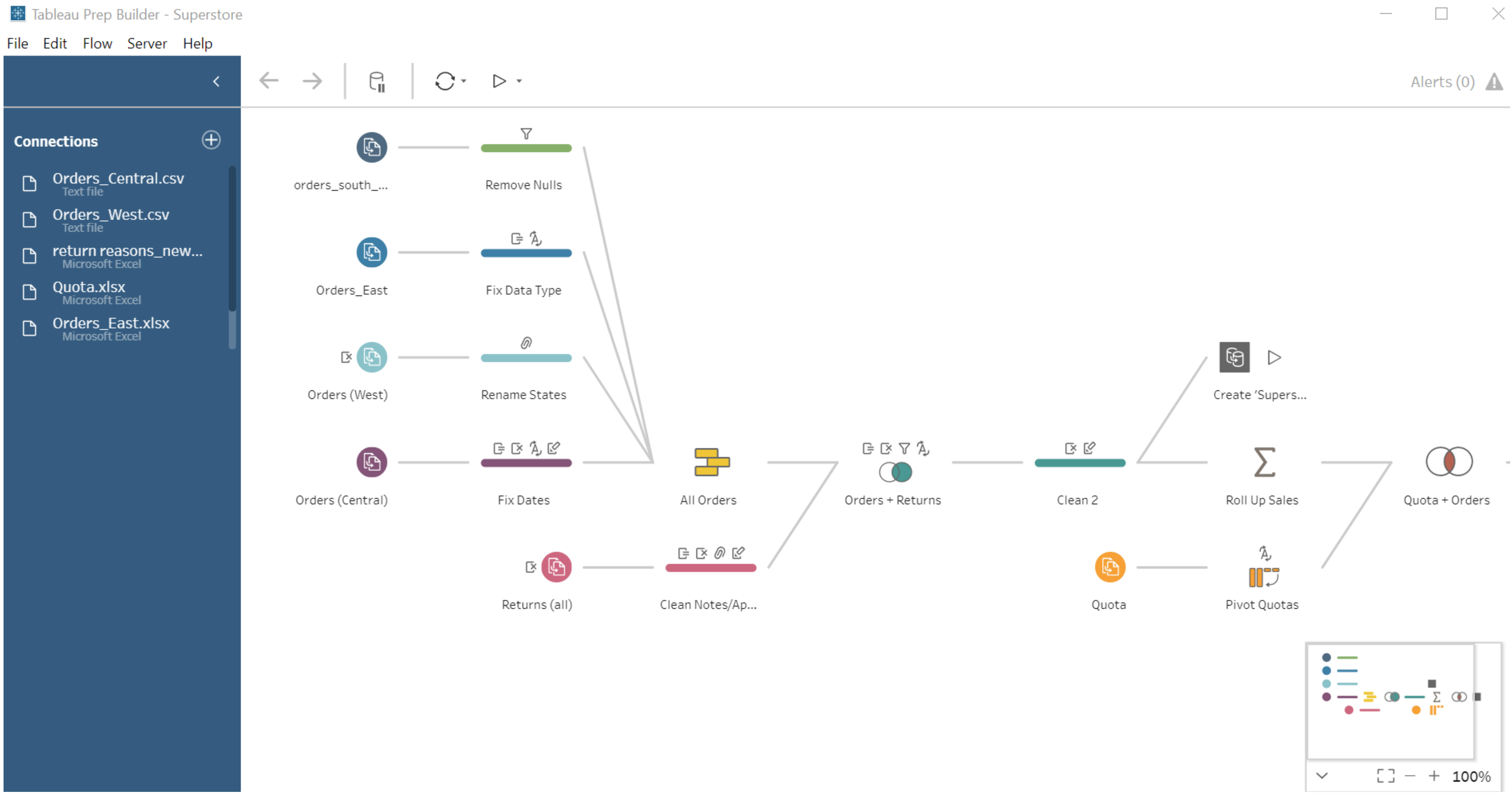
Blue versus green fields

Tableau represents data differently in the view depending on whether the field is discrete (blue), or continuous (green). *Continuous* and *discrete* are mathematical terms. Continuous means "forming an unbroken whole, without interruption"; discrete means "individually separate and distinct."

•Green measures SUM(Profit) and dimensions YEAR(Order Date) are continuous. Continuous field values are treated as an infinite range. Generally, continuous fields add axes to the view.

•Blue measures SUM(Profit) and dimensions Product Name are discrete. Discrete values are treated as finite. Generally, discrete fields add headers to the view.

Tableau also provides Tableau Prep Builder – for cleaning and preparing data sets



Let’s explore a bit

Download the datasets from Canvas; the file is called iris.zip
It’s in the Files area, in a folder called Datasets

sepal_length	sepal_width	petal_length	petal_width	species
4.3	3	1.1	0.1	1
4.4	2.9	1.4	0.2	1
4.4	3	1.3	0.2	1
4.4	3.2	1.3	0.2	1
4.5	2.3	1.3	0.3	1
4.6	3.1	1.5	0.2	1
4.6	3.2	1.4	0.2	1
4.6	3.4	1.4	0.3	1
4.6	3.6	1	0.2	1
4.7	3.2	1.3	0.2	1

Today’s Objectives

Chapter 2 – Data to Insights

- 2.1 Converting Business Problems into Analytics Solutions
- 2.2 Assessing Feasibility
- 2.3 An Analytics Base Table
- 2.4 Features

Descriptive Statistics on a Dataset

Tableau