IDSSP: the International Data Science in Schools Project

Abbreviated Topic Lists

Abbreviated Topic Lists for Unit 1

Note: This document provides an overview of the contents of Unit 1, a curriculum framework for the first year of a two-year pre-calculus course in Data Science, primarily intended for students in their last two years of high school, and for teachers who have to teach the course.

Components in blue denote material relevant to a course to teach teachers.

Topic Area 1.1 Data Science and me

1 What is Data Science?	1.1 Role of data in decision-making — at home, in government, business, industry, sport, ...
	1.2 Introduction to Data Science and the Data Science learning cycle
	1.3 Data Science success stories
	1.4 Data Science disasters
	1.5 Elaboration of steps in the data-science cycle
	1.6T Diverse uses of the term " Data Science "
	1.7T Sources of information about Data Science and its activities
2 What does Data Science have to do with me?	2.1 Data I keep
	2.2 Data about me
	2.3 Data on friends and family
	2.4 What are data?
	2.5 Privacy, security and openness/accessibility — issues and trade-offs
	2.6T Pedagogical issues relating to leading discussions and extracting issues
3 Sources of data	3.1 Examples of data
	3.2 What *are* data?
	3.3 How do we get useful data? Primary versus secondary data.
	3.4 Privacy, security and openness/accessibility -- issues and trade-offs.
	3.5 Thinking critically about data: introduction.
	3.6 Thinking critically about data: data quality and GIGO
	3.7 Thinking critically about data: ways in which data need critical appraisal.
	3.8T Pedagogical issues relating to these topics.
4 Examples of data science problems	4.1 Ideas (with examples) of using data
	4.2 Examples of social and personal consequences
	4.3 How can the prediction process go wrong?
	4.4 Examples of causal explanation and use for control
	4.5 How can the process of finding causes go wrong?
	4.6T Pedagogical issues relating to these topics
5T Extracting pertinent lessons from student discussions (Teachers only)	5.1T Leading discussions and extracting key issues from student contributions
	5.2T Story Telling

Topic Area 1.2 Basic tools for exploration and analysis. Part 1: Tools for a single variable

1 Rectangular data sets	1.1 Observations and variables
	1.2 Numerical versus categorical variables
	1.3 Importing data files in simple formats
	1.4 The need for data cleaning
	1.5T Knowledge of Topic Area 1.6
	1.6T Pedagogical issues relating to these topics
2 Graphics and summaries for a single categorical variable	2.1 Frequency tables and bar charts (counts and proportions)
	2.3 Good ways of ordering groups for displays and tables
	2.4T Proportions versus counts: what works best for what?
	2.5T Weaknesses of pie charts and stacked bar charts
	2.6T Pedagogical issues relating to these topics
3 Graphics and summaries for a single numeric variable	3.1 Graphics: Dot plots and histograms as large-data-set alternative.
	3.2 Features to look out for and their implications
	3.3 Summaries
	3.4 Box plot as plot of summaries
	3.6 Converting numerical variables to categorical: When, why and how? (
	3.6T How the summary measures are obtained
	3.7T Pedagogical issues relating to these topics

Topic Area 1.3 Basic tools for exploration and analysis. Part 2: Pairs of variables

1	**Comparing groups**	1.1 Making comparisons
		1.2 Interpreting "group comparisons"
		1.3 Extension to panel plots
		1.4T Pedagogical issues relating to these topics
		1.5T Comparing and evaluating different presentations
2	**Relationships between two numerical variables**	2.1 Scatter plots
		2.2 Outcome/Response variables versus Predictor/Explanatory variables
		2.3 Construction
		2.4 Structure in scatter plots
		2.5 Basic ideas of prediction
		2.6 Vertical strips as a guide for sketching trend curves by eye
		2.7 How predictions can fail
		2.8 Minimizing average prediction errors
		2.9 Obtaining trend lines and slider-controlled smooths from software
		2.10 (Straight) lines and interpreting the intercept and slope coefficients of a trend line
		2.11 Positive and negative associations
		2.12 Modifications to scatter plots to overcome perceptual problems
		2.13T Working with algebraic expressions
		2.14T Pedagogical issues relating to these topics
3	**Relationships between categorical variables**	3.1 Two-way tables of counts and proportions
		3.2 Side-by-side and separate bar charts or dot charts of proportions
		3.3T Pedagogical issues relating to these topics
4	**Filtering data**	4.1 Filtering data by levels of a categorical variable
		4.2T Pedagogical issues relating to these topics

Topic Area 1.4 **Basic tools for exploration and analysis**. Part 3: Three or more variables

1	**Pairs plots**	1.1 Pairs plots that will cope with categorical as well as numerical variables
		1.2T Pedagogical issues relating to these topics
2	**Subsetting by a third variable**	2.1 Panel plots/facetting and 3-dimensional summary tables
		2.2 Playing or stepping through the sequence of plots in panel display
		2.3 Highlighting subgroups in a scatter plot or dot plot
		2.4T Pedagogical issues relating to these topics.
3	**Other ways of adding information on additional variables to 1- and 2-variable plots**	3.1 Coloring points in dot plots and scatter plots
		3.2 Sizing points in scatter plots
		3.3 Labeling points
		3.4 Strengths and weaknesses of methods of adding information
		3.5T Pedagogical issues relating to these topics
4	**Interactive plots**	4.1 Plots that allow querying of elements.
		4.2 Linked plots, linked plots-and-tables
		4.3T Pedagogical issues relating to these topics.

Topic Area **1.5 Graphs and Tables: how to construct them and when to use them**

1	**Exploration and discovery versus presentation**	1.1 Purposes of a graph; and purposes of a table
		1.2 Infographics and infotables
2	**What makes a graph good or bad?**	2.1 The graphical process: a chain from graph creator to graph interpreter
		2.2 Visualization Principle 1. Use Position along a common scale.
		2.3 Visualization Principle 2. Choose an appropriate Aspect Ratio.
		2.4 Visualization Principle 3. Encoding variables
		2.5 Visualization Principle 4. Supply an informative caption
		2.6 Visualization Principle 5. May need more than one graph
		2.7 Other factors that good software generally gets right by default.
		2.8T Pedagogical issues relating to comparing different types of graphs
3	**What sort of graph should I use?**	3.1 Plotting samples of numerical data to explore relationships
		3.2 Plotting a numerical and a categorical variable
		3.3 Plotting variables that change over time
		3.3 Plotting two numerical variables to explore relationships
		3.5T Pedagogical issues relating to comparing different types of graphs
4	**Tables: their purpose, and how to create good tables**	4.1 Role of tables
		4.2 Principles for making patterns in tabular information
		4.3 When it is often better to use a table rather than a graph
		4.4T Pedagogical issues relating to using tables

Topic Area 1.6 The data-handling pipeline

1 Tool support for the data-handling pipeline	1.1 Automating the data science process
	1.2 Data management principles
	1.3 Case studies of real data-science projects that were done with various toolsets
	1.4T Characteristics and evaluation of some widespread tool-sets
	1.5T Pedagogical issues relating to tool use/mastery
2 Getting and storing data	
	2.1 Data sources
	2.2 Logical data formats
	2.3 Physical file formats
	2.4T Case studies of good data sources
	2.5T Comparison and evaluation of storage approaches
	2.6T Pedagogy issues relating to data sources and storage
3 Tool support for exploring and analysing data	3.1 Automating an analysis
	3.2 Data cleaning
	3.3 Data transformations
	3.4T Learning to use more aspects of the programming language
	3.5T Pedagogical issues for coding
4 Generating presentations the data	4.1 Principles of communication
	4.2 Customizing graphs and tables
	4.3 Combining explanation with graphs/tables
	4.4T Comparison and evaluation of tools that allow generating presentations
	4.5T Pedagogy issues relating to generating presentations

Topic Area 1.7 Avoiding being misled by data

1	**GIGO - "Garbage In, Garbage Out"**	1.1	What do we mean by GIGO?
		1.2	Examples of "garbage". How can we avoid collecting or using garbage?
		1.3T	Pedagogical issues relating to these topics
2	**Bias and what we can do about it**	2.1	Biases due to measurement issues
		2.2	Biases due to selection or filtering in data streams
		2.3	(Discussion Topic) Extrapolating from the data we have to a larger setting
		2.4T	Pedagogical issues relating to these topics
3	**Problems and solutions in reaching causal conclusions**	3.1	Allowing for an important third variable
		3.2	Difference between an observational study and a randomized experiment
		3.3	(Discussion Topic) Extrapolating from the data at hand to a larger setting
		3.4T	Pedagogical issues relating to these topics
4	**Questions that can and cannot be answered by data**	4.1	Learning to ask questions that can be answered from data
		4.2	Learning to spot questions that cannot be answered from the available data
		4.3T	Pedagogical issues relating to these topics
5	**Sampling errors and confidence intervals**	5.1	Random sampling is not perfect
		5.2	Unpacking "the likely extent of sampling error"
		5.3	Experiencing how confidence intervals can be constructed
		5.4T	Pedagogical issues relating to these topics
6	**Randomization variation in experiments**	6.1	Randomized assignment is not perfect
		6.2	(Discussion) When to conclude that observed group differences are real?
		6.3	Experiencing a two-group randomization test
		6.4T	Pedagogical issues relating to these topics

Abbreviated Topic Lists for Unit 2

Note: This document provides an overview of the contents of Unit 2, a curriculum framework for the second year of a two-year pre-calculus course in Data Science, primarily intended for students in their last two years of high school, and for teachers who have to teach the course.

Components in blue denote material relevant to a course to teach teachers.

Topic Area 2.1 Time Series data

1	**Problem elicitation and formulation: Time Series data**	1.1	The nature of time series data
		1.2	Reasons why people are often interested in time series data
2	**Getting the data**	2.1	Obtaining time-series datasets
		2.2	Some common date-and-time variable formats
		2.3	Transforming and reshaping data sets
3	**Exploring the data**	3.1	Basic time-series plots and recognizing features
		3.2	Identifying the trend + seasonal oscillation components
		3.3	Decomposition into trend + season + residual
		3.4	Comparing related series
4	**Analyzing the data: Modelling and Forecasting**	4.1	Forecasting as projecting patterns from the past
		4.2	Making an informal forecast
		4.3	Experience with using a formal forecasting method
		4.4T	Pedagogical issues relating to these topics
5T	**Communicating the Results; next question?**	5.1	Selecting features to be communicated
		5.2	Choosing a communication method
		5.3	Telling the story

Topic Area 2.2 Map data

1	**What are the purposes of Maps?**	1.1	Ubiquitous nature of maps
		1.2	Constituent components of maps
		1.3	Separation of data from display
		1.4	Multiple dimensions
		1.5	Interactive example as a tool for exploratory learning
2	**How do we build and work with location maps?**	2.1	Location and Region Data as common archetypes
		2.2	Plotting points on downloaded map tiles, relationship to scatterplots
		2.3	Coding added-variable-information at location points; interpretation
		2.4	Subsetting/Faceting; ways of showing changes over time
		2.5	Interactivity with location map-plots
3	**How do we build and work with regional maps?**	3.1	Shape files and choropleth maps; region labels
		3.2	Matching regions in a dataset to regions in a shape file
		3.3	Representing two or more variables; issues of scales
		3.4	Perceptual problems with choropleth maps; alternative representations
		3.5	Subsetting/Faceting as a tool; ways of showing changes over time
		3.6	Interactivity with regional maps
		3.7T	Subtleties of color and scale choice – communication enhancement
		3.8T	Distortions and bias – avoiding misleading figures – projections
4	**(TEACHER-only TOPIC) What is a Map, and how is this Data?**	3.3T	Maps as visualization versus maps as data
		3.4T	Maps as data
		3.5T	Finding patterns in data through maps
		3.6T	Overlays on maps

Topic Area 2.3 Text data

1	**Problem elicitation and formulation: Text data**	1.1 Examples of questions to be addressed using natural language
		1.2 Important features of text data
		1.3 Extracting tokens
		1.4 Removing stop words and performing stemming
		1.5T Characteristics of text data
2	**Bag of words analysis of text data**	2.1 Constructing frequency tables of tokens
		2.2 Generating bar charts and word clouds of token frequencies
		2.3 Limitations of unigrams; extracting bigrams
		2.4 Summarizing bigrams; comparing unigrams and bigrams
		2.5 Exploring differences between documents
		2.6 Distinguishing the content of documents
		2.7 Limitations of bag of words analysis
		2.8T Pedagogical issues relating to text data analysis
3	**Sentiment Analysis**	3.1 What is sentiment and why do people want to use it?
		3.2 Merging tokens with sentiment data tables
		3.3 Summarizing sentiment in a document; differences between sources
		3.4 Limitations of sentiment analysis
		3.5T Issues relating to sentiment analysis

Topic Area 2.4 Supervised learning

1 Problem elicitation and formulation: Supervised Classification	1.1 What is classification?
	1.2 Classification models and rules
	1.3 Measuring how well the classification model works
2 Introduction to Classification Trees	2.1 Components of a classification tree and how does it work
	2.2 Misclassification rate
	2.3 Node/leave "purity"
	2.4 Generating Classification and Regression Trees
	2.5 Consequences of misclassification?
3 Growing Classification Trees	3.1 Introduction to R/Python commands to grow and visualize CART
	3.2 When to stop growing the tree
	3.3 What is overfitting? Validation data
	3.4 Pruning
4 Communicating the Results; next question?	4.1 What does the tree tell you about how classifications are made?
	4.2 Using the tree to make decisions
	4.3T Pedagogical issues relating to these topics
5 Introduction to Regression Trees	5.1 Measuring quality of prediction
	5.2 Interpreting regression trees
	5.3 Building regression trees with one predictor variable
	5.4 Building regression trees with more than one predictor variable
	5.5 Comparing trees with a validation set

Topic Area 2.5 Unsupervised Learning

1	**Problem elicitation and formulation: Unsupervised Learning**	1.1 What is unsupervised learning?
		1.2 Creating clusters (groups) of data points based on attributes of the data
		1.3 Contrast with classification or supervised learning
2	**Getting and exploring data**	2.1 Obtain appropriate datasets
		2.2 Interpreting the structure of the data
		2.3 Contrast with Supervised Learning
		2.4 Motivation for identifying clusters automatically
3	**Example of Unsupervised learning algorithm: K-means clustering**	3.1 K-means is a clustering algorithm which is iterative in nature
		3.2 Use of distance metric to assign each data point to a cluster automatically
		3.3 Explanation of iterative procedure
		3.4 The need to repeat with different initial guesses for cluster centers
		3.5 Use a small set of data points to explain the K-means algorithm manually
		3.6 What is an outlier in this context?
		3.7 Discuss distance of an object from the center of its assigned cluster
4	**Implementing K-means clustering on a large data set**	4.1 Introduction to format of the data set.
		4.2 Introduction to the programming environment to use for clustering
		4.3 Clean and transform the data set to observations *versus* features
		4.4 Choose K and run the algorithm using the code snippet provided
		4.5 Interpret the results
		4.6 Change the value for K, and repeat; compare how selections have changed
		4.7T Pedagogical issues relating to these topics
5	**Use in Problem solving**	5.1 When is unsupervised learning as best approach for problem at hand?
		5.2 Exploratory analysis and graphs to summarize the input data.
		5.3 Visualizations to show differences for different values of K
		5.4 Making an optimal choice of K
		5.5 Descriptive statistics for the features in each cluster
		5.6 Interpreting and communicating the results
		5.7 Effect of human factors
6	**Other unsupervised learning methods – Alternatives to K-means clustering**	6.1 Examples when distance-based methods may not be appropriate
		6.2 Motivate need for other clustering methods
		6.3 Visualizations of different cluster shapes unsuited to K-means
		6.4 Examples of visualizations

Topic Area 2.6 Recommender Systems

1	**Problem elicitation and formulation: Recommendation Systems**	1.1 Examples of some recommendation systems
		1.2 Desirable features of recommendation systems
		1.3 Ethical issues for recommendation and personalized systems
		1.4 Additional complexities
		1.5 Communicating the recommendations,
		1.6T Characteristics of a wide variety of recommendation systems
		1.7T Pedagogical issues relating to recommendation systems
2	**The data used by Recommendation Systems**	2.1 Ratings (on user-item pairs): sparsity of the data.
		2.2 Data quality issues
		2.3 Feature data on items, demographic data on users
		2.4 Ethics with data
		2.5 Storing the data
		2.6T Tools to collect and manipulate ratings data
		2.7T Pedagogical issues relating to data for recommendation system
3	**Content-based recommendation**	3.1 Concept: Recommendation based on single-user data, from item similarity
		3.2 Measures of item similarity
		3.3 Analysis and recommendation based on calculating nearest neighbors
		3.4 Analysis and recommendation based on forming clusters of items
		3.5T Tools for calculating similarity, clusters etc
		3.6T Pedagogical issues relating to similarity and clustering
4	**Collaborative-filtering**	4.1 Concept: recommend items that are liked by similar users
		4.2 Define similarity of users
		4.3 Use regression to predict unseen rating from known ones
		4.4 Ethics issues
		4.5T Tools that calculate predicted ratings
		4.6T Pedagogical issues relating to collaborative filter
5	**Evaluation of a Recommendation System**	5.1 User satisfaction; proxies to measure this
		5.2 Measures from information retrieval
		5.3 Ethics rules that apply to user studies
		5.4T Tools to calculate IR measures
		5.5T Pedagogical issues relating to recommendation evaluation

Topic Area 2.7 Interactive Visualization

1 Why visualization? The role of Visualization in the Data Science Learning cycle	1.1 The power of visualization
	1.2 Visual Exploratory Data Analysis
	1.3 Visual Communication and Presentation
	1.4 Considerations and challenges of visualization design
2 Data Types and Visual Variables	
	2.1 Brief introduction to perceptual and cognitive capacity.
	2.2 Hierarchy of visual variables
	2.3 Color
	2.4 Motion
3 Interaction	3.1 The role of interaction
	3.2 Types of interaction – 1. Basic
	3.3 Types of interaction – 2. manipulating the layout
	3.4 Types of interaction – 3. manipulating the data
4 Critique	4.1 Understanding the audience and the task
	4.2 Perceptual biases that affect visualization efficacy
	4.3 Inappropriate encodings of data
	4.4 Scales, legends, decorations

Topic Area 2.8 Confidence intervals and the bootstrap

1 Parameters versus estimates	1.1 Motivation and examples
2 Sampling error	2.1 Behavior of sampling errors in various contexts
	2.2 What is a standard error?
	2.3 Relation between sample size and standard error
	2.4 Inverse-root-n relationship between sample size and sampling error
	2.5 Effect of population size on sampling error
	2.6T Simulation and simulation tools
3 Confidence intervals and their implementation using bootstrapping	3.1 The concept of a confidence interval
	3.2 Interpretation of confidence intervals for a single parameter
	3.3 Experiencing different forms of bootstrap intervals
	3.4 Comparing bootstrap standard error with usual approach
	3.5T Ways to facilitate learning from simulation
4 Investigating performance	4.1 Coverage properties
	4.2 (Optional) More advanced situations
	4.3T Ways to facilitate learning from simulation
5 Differences and ratios	5.1 Constructing and interpreting bootstrap confidence intervals for differences
	5.2T Ways to facilitate learning from simulation
6 Further exploration	6.1 (Optional) CIs when sampling from a set of theoretical distributions
	6.2T Ways to facilitate learning from simulation

Topic Area 2.9 Randomization tests and Significance testing

1 Randomization variation	1.1 The issue of assessing a possible treatment effect
2 Towards the randomization test	2.1 Generation and display of randomization variation
	2.2 Discussion: Concluding that you had evidence of a real difference …
	2.3 Motivate concept of re-randomization
	2.4 Generating the randomization distribution
	2.5 Discussion: scope of inference justified by the experiment
	2.6T Simulation and simulation tools
3 Randomization test	3.1 A general procedure for conducting randomization tests
	3.2 Analyzing data from several experiments and conclusions
	3.3 (*Optional*) Language and idea of P-value and "sidedness" of a test
	3.4 *Optional*) Generalize to 3 or more groups using simple distance measure
	3.5T Simulation and simulation tools
4 (*Optional*) **Investigating the performance of randomization tests in a sampling context**	4.1 Exploring performance of randomization tests under random sampling
	4.2T Simulation and simulation tools
5 Confidence intervals	5.1 (*Optional*) Exploring the performance of confidence intervals from randomized experiments
	5.2T Simulation and simulation tools

Topic Area 2.10 Image data

1 What is Image Data and what sorts of problems does it pose?	1.1 Sources and types
	1.2 Images as data and methods for representing them
2 Some areas of application	2.1 From the world of Art, Classification
	2.2 From Forensic Science, Classification
	2.3T Framing the challenges