# Learning Objectives for Data Concept and Visualization
# (For Exams Beginning October 15, 2020)

## Assignments 7, 9, 11, 13, and 14 include changes from previous syllabus

### Assignment 1: Data Quality

| MODULE TITLE | LEARNING OBJECTIVES |
|---|---|
| **Concept and Impact of Data Quality** | • Summarize concepts of data quality. Understand and describe the impact of data on actuarial work and projects. |
| **Data Quality Principles** | • Understand the categories of data quality principles. Given a principle of data quality, provide an example that illustrates the principle. Understand what is involved in a review of data. |
| **Data Governance** | • Concepts, roles and responsibilities, committees, tools |
| **Data Documentation: Metadata Terminology** | • Describe these aspects of data documentation:<br>➢ Data and metadata terminology<br>➢ Relationship between data documentation and data governance<br>➢ Types and uses of metadata for data scientists |
| **Aggregate Insurance and Statistical Data** | • Explain the regulator and business needs for statistical data |
| **Range of Weight: 0-8 %** | |

### Assignment 2: Sources of Data

| MODULE TITLE | LEARNING OBJECTIVES |
|---|---|
| **Data Science and Data Scientists** | • Understand the fundamental concepts of data science.<br>• Understand different types of data scientists.<br>• Summarize the insurers' use of predictive analytics and data science and the roles of data scientists and data science team. |
| **Insurer Data-Driven Decision Making** | Explain how insurers and risk managers use data-driven decision making |

| | |
|---|---|
| **Insurer Operational Data** | • Understand what typical attributes are made available in each of the following data sources:<br>➢ Policy and Premium Data<br>➢ Claims Information<br>➢ Claim Notes<br>➢ Billing Information<br>➢ Producer Information<br><br>• Understand how corrections for each of these attributes are recorded. |

| | |
|---|---|
| | • Understand timing of collection and updating. Understand how quality can change over time. |
| **The Value of Statistical Plan Data** | • Understand why insurance companies produce statistical files, typical attributes in stat files and the advantages and disadvantages of using stat files over operational data. |
| **Statistical Plans** | • Describe the two basic types of statistical plans and their use in insurers' predictive modeling and ratemaking |
| **Insurance Date Fields and Amount Fields in Statistical Plans** | • Explain the role of the statistical agent and how the date and amount fields are used in the statistical plans. |
| **Classification or Rating Variable Fields and Exposure Data Elements in Statistical Plans** | • Describe the following statistical plan elements by line of business.<br>➢ Classification and rating variable fields<br>➢ Exposure |
| **External Sources of Noninsurance Data** | • Understand how to access and the uses of external sources of data including:<br>➢ Demographic information<br>➢ Customer Financial Information<br>➢ Business Financial Information<br>➢ Behavioral data<br>➢ Driving Records and Motor Vehicle Reports<br>➢ Government sources<br>• Understand who collects the information, for what purposes, how frequently it is updated and how it is distributed. For each of these sources understand what typical attributes are made available.<br>• Understand various derived attributes. Understand if there is a clear way to merge the data into databases used for analysis. |
| **External Sources of Data – Insurance Specific** | • Understand External sources of Insurance Specific data such as:<br>➢ Historical claims reports<br>➢ Industry trend factors<br>➢ Loss Development factors<br>➢ Data available in NCCI Statistical Bulletin |
| **External Data Sources: United States Census Data** | • Understand the two basic formats into which data from the United States Census bureau is organized. |
| **Range of Weight: 9-18 %** | |

## Assignment 3:  Accessing Raw Data

| MODULE TITLE | LEARNING OBJECTIVES |
| --- | --- |
| **Data Classifications** | • Understand the broad classifications of data: quantitative vs qualitative; nominal, ordinal, interval continuous; transactional, snapshots, aggregated.<br>• Understand transaction and snapshot data and how to combine snapshots from various times into transaction files. |
| **Working with Structured Data** | • Understand the various data types: Numeric, string, date, geographic.<br>• Understand how to read, store and display each data type. Understand issues with Date types, such as different formatting across different data sources-coding and why it is used.<br>• Understand structured versus unstructured data and the various forms of unstructured data such as document, map, voice, and image. |
| **Unicode Basics** | • Explain the purpose and functions of Unicode. |
| **Working with Unstructured Data** | • Be aware of other external sources of data such as:<br>➢ Social media feeds<br>➢ Web APIs<br>➢ ACID<br>➢ NoSQL |
| **Text File Formats** | • Explain how to read and write data to delimited and fixed text file formats |
| **Dataframes** | • Explain how data can be entered and stored in these types of dataframes.<br>➢ Relational databases<br>➢ Data warehouses<br>➢ Excel<br>➢ R<br>➢ Python |
| **Data Exchange** | • Describe how the exchange of data is facilitated by:<br>➢ HTML<br>➢ XML<br>➢ JSON<br>➢ Data marts |
| **Obtaining Data from the Internet** | • Describe methods for obtaining data from the internet |

| | |
|---|---|
| **Data Profiling** | • Demonstrate ability to profile data including: inspecting rows of data, read data catalogs and metadata. Create descriptive statistics and graphs that profile the data. |
| **Messy Data** | • Detect and remediate missing, miscoded and anomalous data. Understand sampling bias and clustering of values. |
| **Testing Data** | • Work with small test data and create sample data using simple filters. Be able to sample from related tables. |
| **Range of Weight: 9-18 %** | |

### Assignment 4: Working with Data

| MODULE TITLE | LEARNING OBJECTIVES |
|---|---|
| **Querying Data** | • Explain how to query data from a database using Structured Query Language<br>➢ Querying Data<br>➢ SELECT, FROM, and WHERE Statements<br>➢ Retrieving Columns<br>➢ Retrieving Distinct Data<br>➢ Aggregate Functions<br>➢ Grouping Data |
| **Joining Data Tables** | • Explain how to join data tables using Structured Query Language<br>➢ Multiple Tables<br>➢ Joining Two Tables<br>➢ Joining More Than Two Tables<br>➢ Subqueries |
| **Advanced SQL Topics** | • Describe these issues in Structured Query Language: Indexes; Null values; User-defined functions; Large-data access<br>➢ Indexes<br>➢ Null Values<br>➢ User-Defined Functions<br>➢ Large-Data Access |
| **String Functions** | • Understand the use of string functions in common data preprocessing software.<br>➢ String Processing With SQL<br>➢ String Processing With Regular Expressions |

| | |
|---|---|
| **Working With Regular Expressions in String Processing** | • Understand the use of common regular expressions for pattern matching.<br>➢ Basic Regular Expressions<br>➢ Application of Regular Expressions |
| **Using Hash Tag Functions** | • Explain how to use hash functions with databases.<br>➢ Common Hash Functions<br>➢ Hash Functions for Equality Testing<br>➢ Hash Functions for Hash Tables<br>➢ Hash Functions for Data Segmentation |
| **Insurance Applications of Data Preparation** | • Apply Structured Query Language to develop profiles from premium and loss data summaries.<br>➢ Summarize an Earned Premium Table to a Policy-Term Summary<br>➢ Summarize a Policy-Level Loss File to a Book-Level Summary<br>➢ Join the Policy Table to the Loss Table<br>➢ Join a Table Showing Demographic Information to the Premium and Loss Tables<br>➢ Create Profiles Using the Information From All Tables |
| **Range of weight: 0-8 %** | |

### Assignment 5: Regulations and Privacy Issues

| MODULE TITLE | LEARNING OBJECTIVES |
|---|---|
| **Data Regulation** | • Describe the fundamental concepts associated with government data regulation.<br>➢ U.S. Data Regulations<br>➢ U.S. State Laws<br>➢ International Legislation |
| **Range of weight: 0-8 %** | |

## Assignment 6: Data Tools and Exploratory Visualization

| MODULE TITLE | LEARNING OBJECTIVES |
|---|---|
| **Exploratory Data Analysis and Data Transformation** | • Summarize these aspects of exploratory data analysis (EDA): Uses of EDA; Role of metadata in EDA; Data transformations identified through EDA<br>➢ Exploratory Data Analysis<br>➢ Data Transformation |
| **Characterizing Data with Univariate Displays** | • Describe the use of univariate descriptive statistics and displays, and some basic techniques for meaningful data characterization.<br>➢ Basic Descriptive Statistics<br>➢ Graphs, Tables, and Charts<br>➢ Displaying and Assessing Time Series Data<br>➢ Bucketing for Categorical Variables |
| **Identifying and Treating Data Anomalies** | • Describe these aspects of data anomalies: Types of data anomalies; Methods to detect anomalies; Adjustments to data to reduce the impact on analysis<br>➢ Defining Types of Data Anomalies<br>➢ Detecting Data Anomalies<br>➢ Adjustments to Data |
| **Using Multivariate Summaries and Displays** | • Describe the use of multivariate summaries and displays to analyze data, detect outliers, and/or formulate preliminary hypotheses.<br>➢ Pivot Tables<br>➢ Contingency Tables<br>➢ Linear and Nonlinear Correlations<br>➢ Scatterplots and Correlations<br>➢ Heat Maps |
| **Visualization Methods** | • Describe these aspects of data visualization: Data preparation for visualization; Basic concepts and methods of data visualization<br>➢ Data Preparation for Visualization<br>➢ Basic Concepts and Methods of Data Visualization |
| **Data Visualization Displays** | • Explain how to visualize data using various types of displays.<br>➢ Tables<br>➢ Dashboards<br>➢ Charts and Graphs<br>➢ Maps |
| **Range of weight: 0-9 %** | |

# External Readings – Exploratory Data Analysis

### Assignment 7: *Exploratory Data Analysis with R* - Roger Peng

| TOPICS | LEARNING OBJECTIVES |
|---|---|
| **Chapter 3,4 – Introduction to R and Managing Data Frames** | • Introduction<br>• Choose functions in the dplyr function to manage data frames and to apply basic data screening procedures |
| **Chapter 5 – Exploratory Data Analysis Checklist** | • Construct an exploratory data analysis using the 10-point checklist |
| **Chapter 6,7 – Graphics and Exploratory Graphs** | • Apply fundamental principles of analytic graphics<br>• Develop simple summaries and exploratory graphs that optimize data visualization at the beginning stages of data analysis |
| **Chapter 8-16 - Plotting Systems** | ompare the base, lattice, and ggplot2 plotting systems<br>reate plots on graphic devices including screens and files<br>reate graphics on the screen device in the base plotting ystem, including annotation, regression lines and multiple lots<br>• Design plots in R that optimize the use of color and transparency, including applications to large datasets<br>• Apply the qplot function in ggplot2 for elegant expression of plot components<br>• Apply the key components of a ggplot2 plot to build plots in layers, including data frames, aesthetic mappings, geoms, facets, stats, scales, and coordinate system |
| **Chapter 17 - Data Analysis Case Study: Changes in Fine Particle Air Pollution in the US** | • Conduct an exploratory data analysis in R using raw data |
| **Range of weight: 9-18 %** | |

### Assignment 8: *Visualizing Data* - William Cleveland

| TOPICS | LEARNING OBJECTIVES |
|---|---|
| **Chapter 2 - Univariate Data** | • Summarize the characteristics and purposes of visualizing univariate data<br>• Construct Q-Q plots and normal plots to evaluate data distributions<br>• Develop fits and residuals, spread of fits and residuals, |

| | |
|---|---|
| | quantile plots of residuals and spread location plot<br>• Evaluate when to use log scale for displaying data<br>• Apply transformations of non-normal variables to approximate normality<br>• Apply robust estimation to mitigate distortion from outliers in the data |
| **Range of weight: 0-8 %** | |

**Assignment 9: *Data Visualization* - Kieran Healy**

| | |
|---|---|
| **Chapter 1,2 – Introduction** | • Explain how well-designed quantitative tables and graphs communicate important information effectively<br>• Compare the types of data and relationships included in quantitative stories<br>• Distinguish these summarization measures: average, variation, correlation, ratio |
| **Chapter 1 – What Makes Bad Graphs Bad** | • Explain why these types of graphs don't communicate effectively: 3-D charts, stacked bar, unusual infographics<br>• Explain the three types of badness and describe examples of each |
| **Chapter 1.3- Visual Perception and Graphical Communication** | • Explain how the eye and brain process signals in the mechanics of sight<br>• Illustrate how the attributes of form, color, and spatial position affect preattentive processing<br>• Apply preattentive attributes to design tables and graphs with optimal visual emphasis<br>• Summarize the Gestalt principles of visual perception and how we group objects in particular ways |
| **Chapter 3- Fundamental of Making GraphsGraphs** | • Explain what tidy data is and how it related to making graphs<br>• Explain how to build plots by layer<br>• Explain aesthetics, geoms, and mapping in ggplot and understand how to implement in ggplot |

| | |
|---|---|
| **Chapter 4, 5 – Show the right numbers with ggplot** | • Explain how communication-oriented design supports the objectives of highlighting and organizing information<br>• Understand the concept of the grammar of graphics<br>• Be able to apply aesthetics and geoms<br>• Understand developing graphs by layer<br>• Understand how and why transform data for graphing with dplyr<br>• Design these secondary data components for optimal communication in graphs: trend lines, reference lines, annotations, scales, tick marks, grid lines, legends<br>• Recommend graph design strategies to address presentation of multiple variables: multiple units of measure in a single graph, combining multiple graphs in a series |
| **Chapter 6 through 6.4** | • Understand how to present model results<br>• Understand how to compare and assess models<br>• Understand how to use model objects<br>• Understand how to generate prediction graphs |
| **Chapter 8 - Refine your Plot** | • Apply more advanced principles and techniques in order to create a compelling statistical narrative: color, text, themes, redrawing bad slides, |
| **Chapter 7 – Draw Maps** | • Explain how to map US State level data<br>• Explain when other graphs may be preferable to maps<br>• Understand Ur-choropleths<br>• Apply Small-Multiple maps<br>• |
| **Range of weight: 8-16 %** | |

### Assignment 10: *"Infovis and Statistical Graphics: Different Goals, Different Looks"*, Andrew Gellman and Antony Unwin

| | |
|---|---|
| **Infovis and Statistical Graphics** | • Explain how the different goals of Statistical visualization and Information graphics affect design of graphs<br>• Distinguish examples of flaws and strengths from Visualization projects that received praise<br>• Describe an example of a graph that displays benefits of both statistical graphs and Infoviz |
| **Range of weight: 0-8 %** | |

### Assignment 11: Tables vs. Graphs – Articles by Stephen Few

| | |
|---|---|
| **Tables vs. Graphs, Table Design** | • Evaluate the features of tables and graphs for their optimal use scenarios<br>• Show how the thoughtful design of information and support components lead to clear and efficient communication with tables |
| **Range of weight: 0-8%** | |

### Assignment 12: Insurance Applications

| MODULE TITLE | LEARNING OBJECTIVES |
|---|---|
| **Fundamentals of Modeling Data for Insurance Applications** | • Describe the fundamental applications of data mining statistical tools for data preparation.<br>➢ Overview of Potential Modeling Methods<br>➢ Screening of Variables for Inclusion in the Final Modeling Dataset<br>➢ Transformation of Variables, Including Binning<br>➢ Identification and Treatment of Missing Values<br>➢ Identification of Outliers and Errors<br>➢ Separation of the Data Into Training and Holdout Data<br>➢ Testing of Models |

| | |
|---|---|
| **Creating Datasets for Underwriting Models** | • Describe the fundamental concepts and challenges associated with preparing modeling data for underwriting applications.<br>➢ Identify Target Variable<br>➢ Make Insurance Adjustments<br>➢ Identify Data Sources<br>➢ Gather Metadata<br>➢ Join Data From Different Sources<br>➢ Perform Exploratory Data Analysis and Data Cleansing<br>➢ Transform Variables<br>➢ Identify Potential Independent Variables<br>➢ Perform Data Reduction<br>➢ Create Training/Test/Holdout Samples<br>➢ Document Data Preparation Work |
| **Creating Datasets for Claims Models** | • Describe the fundamental concepts and challenges associated with preparing modeling data for claims applications.<br>➢ Data for Claims Triage Modeling<br>➢ Data for Claims Fraud Modeling<br>➢ Data for Next-Best-Action Modeling |
| **Range of weight: 0-8 %** | |

**Assignment 13: *Planning a Modeling Project*, CRISP-DM, Harvard Business Review**

| MODULE TITLE | LEARNING OBJECTIVES |
|---|---|
| **CRISP-DM** | • Apply the CRISP-DM framework to the process of planning out a predictive analytics project. |
| **Algorithms Need Manages, Too** | • Understand the limitations of modeling and apply this to determine how to leverage models effectively, and how to identify useful input data. |
| **The Discipline of Business Experimentation** | • Understand the differences between working with observational data and experimental data, and why observational data can lead one astray.<br>• Identify methods that can help the modeler make better use of observational data.<br>➢ Understand and evaluate the role of business experiments in model implementation. |
| **Range of weight: 4-8 %** | |

## Assignment 14:  Ethical Considerations and Public Perception (5-9%)

| MODULE TITLE | LEARNING OBJECTIVES |
|---|---|
| **O'Neil, Chapters 6 and 9** | • Describe ways models can promote fairness and ways in which they can perpetuate social inequalities.<br>• Recognize concerns (whether fair or not) that are commonly expressed about predictive models in the press and in political discourse.<br>• Recognize the social implication of threshold selection models.<br>• Identify situations where ethical care and oversight would be advisable. |
| **Loukides** | • Describe the key principles of Loukides et al.'s ethical framework Checklist for Data Science projects and the plusses and minuses of checklists. |
| **Wattenberg** | • Recognize the social implication of threshold selection models.<br>• Describe the importance of feedback in model calibration. |

# Complete Text References for Exam 3

**Text references are alphabetized by the Abbreviation column.**

| Citation | Abbreviation | Assignment | Source |
|---|---|---|---|
| Cleveland, W., Visualizing Data, Hobart Press, 1993, ch. 2 | Cleveland | 8 | **B** |
| Chapman, P., et al., *CRISP-DM 1.0*, SPSS, 2000, https://the-modeling-agency.com/crisp-dm.pdf | CRISP-DM | 13 | **O** |
| Few, Stephen, "Designing Effective Tables and Graphs," perceptualedge.com | Few | 11 | **O** |
| Few, Stephen, "Rare Business Asset: Tables and Graphs that Communicate," perceptualedge.com | Rare | 11 | **O** |
| Gellman, A., and Unwin, A., "Infovis and Statistical Graphics: Different Goals, Different Looks," http://www.stat.columbia.edu/~gelman/research/published/vis14.pdf | Gellman | 10 | **O** |
| Healy, K., *Data Visualization*, Princeton University Press, 2019 | Healy | 9 | **B** |
| Loukides, M., Mason, H., and Patil, D., *Ethics and Data Science*, O'Reilly Media, 2018. Free download from amazon.com | Loukides | 14 | **O** |
| Luca, M., Kleinberg, J., and Mullainathan, S., "Algorithms Need Managers, Too", Harvard Business Review, Jan – Feb 2016, https://hbr.org/2016/01/algorithms-need-managers-too | Luka | 13 | **O** |
| O'Neil, C., *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Broadway Books, 2017: Chapters 6 and 9 | O'Neil | 14 | **B *** |
| Peng, R., *Exploratory Data Analysis With R*, Leanpub, 2015 | Peng | 7 | **B** |
| Thomke, S., and Manzi, J., "The Discipline of Business Experimentation", Harvard Business Review, Dec. 2014 https://hbr.org/2014/12/the-discipline-of-business-experimentation | Thomke | 13 | **O** |
| Wattenberg, M., Viégas, F., and Hardt, M., "Attacking Discrimination with Smarter Machine Learning" (https://research.google.com/bigpicture/attacking-discrimination-in-ml/) | Wattenberg | 14 | **O** |

**Source Key**

| | |
|---|---|
| **B** | Book—may be purchased from the publisher or bookstore or borrowed from the CAS Library. |
| **P** | Online Publication |
| ***** | Available as part of the study packet for Exam 3 |