Glendale Unified School District

High School

April 16, 2019

Department:	Mathematics
Course Title:	Introduction to Data Sciences
Course Code:	3260/3261
School(s) Course Offered:	2019-20: Crescenta Valley High School, 2020-21: Glendale High, Clark Magnet High School, Hoover High School
UC/CSU Approved	Y, “c”
Course Credits:	Full Year (10)
Recommended Prerequisite:	Integrated Mathematics I Integrated Mathematics II
Textbook:	Introduction to Data Science curriculum provided by UCLA Center X

Course Overview: The Introduction to Data Science course will emphasize the use of statistics and computation as tools for creative work, as a means of telling stories with data. Its content will prepare students to "read" and think critically about existing data stories. Ultimately, this course will be about how we tell good stories from bad, through a practice that involves compiling evidence from one or more sources and often requires hands-on examination of one or more data sets. It will develop the tools, techniques and principles for reasoning about the world with data. It will present a process that is iterative and authentically inquiry-based, comparing multiple "views" of one or more data sets. Inevitably, these views are the result of some kind of computation, producing numerical summaries or graphical displays. Their interpretation relies on a special kind of computation, simulation, and modeling to describe the uncertainty in each view. This kind of reasoning is exploratory and investigatory, sometimes framed as hypothesis evaluation and sometimes as hypothesis generation. R, the statistical programming language used by academics and industry, will be used to bring data science to life.

The main goal of the Introduction to Data Science course is to teach students to think critically about and with data. This new and innovative curriculum will meet the Common Core State Standards (CCSS) for High School Statistics and Probability, relevant second-year Algebra probability standards, the Modeling standard, relevant mathematics standards. This course emphasizes the CCSS High School — Statistics and Probability Standards that involve the study of data science. Students authentically apply the Standards for Mathematical Practice throughout the course.

Introduction to Data Science will develop the tools, techniques and principles for reasoning about the world with data, with a special emphasis on data collected through participatory sensing, an emergent and important data type encountered in many disciplines, including business, biology, engineering, computer science, and statistics. We will present a pedagogical process that is iterative and authentically inquiry-based, and this student-based inquiry will lead to comparing multiple "views" of one or more data sets. Inevitably, these views are the result of some kind of computation, producing numerical summaries or graphical displays. Their interpretation relies on a special kind of computation – simulation– to model the uncertainty in each view.

The use of participatory sensing data will put data collection into the hands of students and, as a consequence, students will function as researchers making truly original discoveries about the real world. Students will learn to generate hypotheses, to fit statistical and mathematical models to data, to implement these models algorithmically, and to evaluate how well these models fit reality. Our course will rely on R, an open-source programming language has long been the standard for academic statisticians and analysts in industry. Through R, students will learn to compute with data to develop graphical and numerical summaries to both communicate findings and to generate further exploration.

This course is an introduction to the practice of data science: reasoning about the world with data. The course applies concepts from statistics and probability, alongside computation and visualization, as a means of processing data to learn about the world. An emerging academic discipline, data science creates a basis for thinking about and with data and understanding the ways in which data operate to shape our world. There are four goals encompassing this course divided into four units:

Unit 1: Data are all around us.
Unit 2: Making inferences using models and plots.
Unit 3: Understand data sources, special data structures, and modes of data collection.
Unit 4: Making inferences using randomization and simulation.

Course content:

Semester A

Unit 1 *(approximately 5 weeks)*

A. *Content Standards:* S-ID 1, S-ID 6, S-ID 5, S-ID 2, S-ID 3

This unit will introduce the idea of "data," fundamental to the rest of the course. Traditional statistics courses consist of understanding data from only a small subset of data generation processes, namely those collected through random sampling or random assignment in scientific experiments. This unit exposes students to a wider world of data, and will help students see how to make sense of these ubiquitous data types. This unit will motivate the idea that data and data products (charts, graphs, statistics) can be analyzed and evaluated just like other arguments, such as those used by journalists. Students will begin learning how to construct multiple views of data in an attempt to uncover new insights about the world, using the techniques of descriptive statistics. This will require the introduction of the computational tool R, through the interface of RStudio. Standard graphical displays, like histograms and scatterplots, will be introduced in RStudio, as well as measures of center and spread.

B. Unit Assignment(s):
Assignment: Lesson 14: Variables, Variables, Variables Objective:
Students will learn how to read and interpret multiple variable plots: bivariate scatterplots, multiple variable scatterplots stacked bar plots and side-by-side bar plots. They will summarize their learning about multiple variable plots using a four-fold graphic organizer. This assignment is followed by 2 labs in which student use their skills in interpreting multi-variable plots to answer real-world questions such as; Do healthier snacks cost more or less than less healthy snacks? And what other variables seem to be related to the cost of a snack?
Describe their relationships.

Unit 2 *(approximately 5 weeks)*
A. *Content Standards:* S-DI 2, S-ID 3, S-ID 4, S-IC 2, S-CP 2, S-CP 9, S-IC 6
This unit deepens the informal reasoning skills developed in Unit 1 by enriching students' technical vocabulary and developing more precise analytical tools. Most importantly, this unit introduces the formal concept of probability as a tool for understanding that sometimes patterns observed in data are not "real." Traditional courses attempt to develop this understanding through the development of abstract mathematical probability concepts, but IDS creates enduring understanding by teaching students to design and implement simulations using pseudo- random number generators. Students will be introduced to linear models - the most common form of modeling in introductory statistics classes - which will serve as the foundation to learn more complex modeling techniques that use the computer technology available to them later in the course, including smoothing techniques and tree-based models.

B. Unit Assignment(s):
Assignment: Lesson 8: How Likely Is It?
Objective:
Students will understand the basic rules of probability. They will learn that previous outcomes do not give information about future outcomes if the events are independent. In this lab, we're going to estimate the probability that a rap song will be chosen from a playlist with both rap and rock songs, if the choice is made at random. The playlist we'll work with has 100 songs: 39 are rap and 61 are rock.

Semester B

Unit 3 *(approximately 5 weeks)*

A. *Content Standards:* S-IC 1, S-IC 3, S-IC 6

Unit 3 focuses on data collection methods, including traditional methods of designed experiments and observational studies and surveys. It introduces students to sampling error and bias, which cause problems in analysis made from survey data. Participatory Sensing is presented as another method of data collection, and students learn to design participatory sensing campaigns that will allow them to address particular statistical questions.

B. Unit Assignment(s):
End of Unit Project
Practicum: What Does Our Campaign Data Say?
Objective:
Students will answer the statistical question they generated at the beginning of the Participatory Sensing campaign creation lesson. They will use RStudio to make graphical representations or numerical summaries of their data to answer their question.

Experiments in the medical field that involve new treatments (new medications) are called clinical trials. You have received a data set that shows the results from Sir Austin Bradford Hill's first randomized study in 1948 examining the effects of the antibiotic Streptomycin on 107 tuberculosis patients. You and a partner will use this data set to find out if Streptomycin is an effective treatment for tuberculosis.
A short article about tuberculosis facts can be found at:
http://www.cdc.gov/tb/publications/factsheets/general/tb.htm
Since this is an experiment, answer the following questions below. You may need to research the answer to some of the questions.

- What is the research question?
- Who are the subjects that participated in the experiment?
- What is the treatment?
- Who is in the treatment group?
- Who is in the control group?
- How were the subjects assigned to each group?
- What population is this experiment representative of?
- What is the variable that we will be measuring?
- What is the outcome of this experiment?

Create a four-to-five-slide, 5-minute presentation that shows your results. Be sure to include detailed explanation of how you and your partner decided to conduct your simulation. Each person must participate in the presentation. In addition to the presentation, submit a two to four page, double-spaced summary of your analysis.

Unit 4 *(approximately 5 weeks)*

A. *Content Standards:* S-IC 2, S-ID 6, S-ID 7, S-ID 8, S-IC 6

This unit will develop modeling skills, beginning with learning to fit and interpret least squares regression lines and learning to use regression to make predictions. Students will learn to evaluate the success of these predictions and so compare models for their predictive accuracy. Modern algorithmic approaches to regression are presented, and students will strengthen algorithmic thinking skills by understanding how and why these algorithms help data scientists make accurate predictions from data. Students engage in a complete modeling experience in which they apply the skills and concepts learned in the previous units.

B. Unit Assignment(s):

Assignment: Lesson 8: What's the Trend? Objective:

Students will understand that the regression line is a model for a linear association (trend). They will learn to identify the direction and strength of trends.

Distribute What's the Trend? (LMR_4.7). Students will analyze the two scatterplots on the handout. The Profits per Explosion plot shows the relationship between the number of explosions in Michael Bay's movies and the profit earned by each movie. The Scores Over Time plot shows the relationship between M. Night Shyamalan movies made since The Sixth Sense was released in 1999 and their Internet Movie Database (IMBD) scores.