El Camino College
COURSE OUTLINE OF RECORD – Approved

I. GENERAL COURSE INFORMATION

Subject and Number:	**Computer Science 8**
Descriptive Title:	**Foundations of Data Science**
Course Disciplines:	**Computer Science**
Division:	**Mathematical Sciences**

Catalog Description:
This course examines the foundations of data science from three perspectives: inferential thinking, computational thinking, and real-world relevance. The course combines programming skills and statistical inference to ask questions and explore problems encountered in real-world datasets, from multiple fields of study, career paths, and everyday life. It also delves into social and legal issues surrounding data analysis, including issues of privacy and data ownership.

Conditions of Enrollment:
Prerequisite: Math 67 or Math 73 or Math 80 with minimum grade of C or appropriate assessment

Course Length:	**X Full Term**	**Other (Specify number of weeks):**
Hours Lecture:	**3.00 hours per week**	**TBA**
Hours Laboratory:	**3.00 hours per week**	**TBA**
Course Units:	**4.00**	

Grading Method:	**Letter**
Credit Status:	**Associate Degree Credit**

Transfer CSU:	**X**	**Effective Date:**
Transfer UC:	**No**	

General Education:

El Camino College:

CSU GE:

IGETC:

II. OUTCOMES AND OBJECTIVES

A. COURSE STUDENT LEARNING OUTCOMES (The course student learning outcomes are listed below, along with a representative assessment method for each. Student learning outcomes are not subject to review, revision or approval by the College Curriculum Committee)

1. **Using Specifications:** Student will write correct computer code that will produce appropriate visualizations and analysis of data.

2. **Tracing the Execution:** Students, when given a code segment, will be able to trace the execution and give the output.

3. **Explaining Concepts:** Students will be able to explain data science concepts related to causality versus correlation, hypothesis testing, prediction and classification.

The above SLOs were the most recent available SLOs at the time of course review. For the most current SLO statements, visit the El Camino College SLO webpage athttp://www.elcamino.edu/academics/slo/.

B. Course Student Learning Objectives (The major learning objective for students enrolled in this course are listed below, along with a representative assessment method for each)

1. Explain and justify conclusions about data and draw robust conclusions based on incomplete information.
   Assessment method: Objective Exams

2. Develop written computer code, demonstrating computational thinking and skills, for visualizing and analyzing data.
   Assessment method: Computer Programming Assignments (other)

3. Make and justify predictions based on machine learning.
   Assessment method: Performance Exams

4. Interpret and communicate data and results using a vast array of real-world examples.
   Assessment method: Written Homework

III. OUTLINE OF SUBJECT MATTER (Topics are detailed enough to enable a qualified instructor to determine the major areas that should be covered as well as ensure consistency from instructor to instructor and semester to semester.)

Lecture or Lab	**Approximate Hours**	**Topic Number**	**Major Topics**
Lecture	1	I	Experiments A. Establishing Causality B. Randomization
Lecture	1.5	II	Programming in Python A. Expressions B. Names C. Call Expressions D. Tables
Lecture	1.5	III	Data Types A. Numbers B. Strings C. String Methods D. Comparisons
Lecture	1	IV	Sequences A. Arrays B. Ranges
Lecture	3	V	Tables A. Building Tables B. Selecting Rows C. Selecting Columns
Lecture	3	VI	Visualization A. Categorical Distributions B. Numerical Distributions C. Overlaid Graphs
Lecture	6	VII	Functions A. Applying Functions to Columns B. Groups C. Joins
Lecture	3	VIII	Randomness A. Conditional Statements B. Iteration C. Simulation D. Finding Probabilities
Lecture	3	IX	Sampling and Empirical Distributions A. Sampling B. Models
Lecture	5	X	Hypotheses Testing A. Comparing Distributions B. Decisions and Uncertainty C. A/B Testing
Lecture	3	XI	Causality A. A/B Testing B. Comparing two Samples
Lecture	3	XII	Estimation A. Percentiles B. Confidence Intervals C. Interpreting Confidence

Lecture	6	XIII	Statistics A. Mean B. Variability C. The Normal Distribution D. Sample Means
Lecture	6	XIV	Prediction A. Correlation B. Linear Regression C. Least Squares D. Residuals
Lecture	2	XV	Regression Inference A. Regression Models B. True Slope Inference C. Prediction Intervals
Lecture	6	XVI	Classification A. Nearest Neighbors B. Classifiers C. Decisions
Lab	1	XVII	Experiments A. Establishing Causality B. Randomization
Lab	1.5	XVIII	Programming in Python A. Expressions B. Names C. Call Expressions D. Tables
Lab	1.5	XIX	Data Types A. Numbers B. Strings C. String Methods D. Comparisons
Lab	1	XX	Sequences A. Arrays B. Ranges
Lab	3	XXI	Tables A. Building Tables B. Selecting Rows C. Selecting Columns
Lab	3	XXII	Visualization A. Categorical Distributions B. Numerical Distributions C. Overlaid Graphs
Lab	6	XXIII	Functions A. Applying Functions to Columns B. Groups C. Joins
Lab	3	XXIV	Randomness A. Conditional Statements B. Iteration C. Simulation D. Finding Probabilities

Lab	3	XXV	Sampling and Empirical Distributions A. Sampling B. Models
Lab	5	XXVI	Hypotheses Testing A. Comparing Distributions B. Decisions and Uncertainty C. A/B Testing
Lab	3	XXVII	Causality A. A/B Testing B. Comparing two Samples
Lab	3	XXVIII	Estimation A. Percentiles B. Confidence Intervals C. Interpreting Confidence
Lab	6	XXIX	Statistics A. Mean B. Variability C. The Normal Distribution D. Sample Means
Lab	6	XXX	Prediction A. Correlation B. Linear Regression C. Least Squares D. Residuals
Lab	2	XXXI	Regression Inference A. Regression Models B. True Slope Inference C. Prediction Intervals
Lab	6	XXXII	Classification A. Nearest Neighbors B. Classifiers C. Decisions
Total Lecture Hours		**54**	
Total Laboratory Hours		**54**	
Total Hours		**108**	

IV. PRIMARY METHODS OF EVALUATION AND SAMPLE ASSIGNMENTS

A. PRIMARY METHOD OF EVALUATION

Problem solving demonstrations (computational or non-computational)

B. TYPICAL ASSIGNMENT USING PRIMARY METHOD OF EVALUATION

Probability and Sampling

When a company produces medical devices, it must be sure that its devices will not fail. Sampling is used ubiquitously in the medical device industry to test how well devices work.

Suppose you work at a company that produces syringes, and you are responsible for ensuring the syringes work well. After studying the manufacturing process for the syringes, you have a hunch that they have a 1% failure rate. That is, you suspect that 1% of the syringes won't work when a doctor uses them to inject a patient with medicine.

To test your hunch, you would like to find at least one faulty syringe. You hire an expert consultant who can test a syringe to check whether it is faulty. But the expert's time is expensive, so you need to avoid checking more syringes than you need to.

Important note: This exercise asks you to compute numbers that are related to probabilities. For all questions, you can calculate your answer using algebra, **or** you can write and run a simulation to compute an approximately-correct answer. (For practice, we suggest trying both.) An answer based on an appropriate simulation will receive full credit. If you simulate, use at least **5,000** trials.

Question 1. Suppose there is indeed a 1% failure rate among all syringes. If you check 20 syringes chosen at random from among all syringes, what is the chance that you find at least 1 faulty syringe? (You may assume that syringes are chosen with replacement from a population in which 1% of syringes are faulty.) Name your answer chance_to_find_syringe.

Question 2. Continue to assume that there really is a 1% failure rate. Find the smallest number of syringes you can check so that you have at least a 50% chance of finding a faulty syringe. (Your answer should be an integer.) Name that number num_required_for_50_percent. **It's okay if your answer is off by as many as 11 for full credit.**

Question 3. A doctor purchased 5 syringes and found 4 of them to be faulty. Assuming that there is indeed a 1% failure rate, what was the probability of **exactly 4** out of 5 syringes being faulty?

Question 4. Assuming that there is indeed a 1% failure rate, assign order to a list of the numbers 1 through 7, ordered by the size of the quantities described below from smallest to largest. For example, order will start with 2 because list item 2 ("Zero") is the smallest quantity.

1. One half
2. Zero
3. The chance that **zero** out of 5 syringes are faulty.
4. The chance that **at least 1** out of 5 syringes is faulty.
5. The chance that **exactly 4** out of 5 syringes are faulty.
6. The chance that **at least 4** out of 5 syringes are faulty.
7. The chance that **all 5** out of 5 syringes are faulty.

C. COLLEGE LEVEL CRITICAL THINKING ASSIGNMENTS

1. Data Visualization

In this project, you'll explore data from Gapminder.org, a website dedicated to providing a fact-based view of the world and how it has changed. That site includes several data visualizations and presentations, but also publishes the raw data that we will use in this project to recreate and extend some of their most famous visualizations.

The Gapminder website collects data from many sources and compiles them into tables that describe many countries around the world. All of the data they aggregate are published in the Systema Globalis. Their goal is "to compile all public statistics; Social, Economic and Environmental; into a comparable total dataset." All data sets in this project are copied directly from the Systema Globalis without any changes.

Using code perform the following operations:

Question 1. Create a table called b_pop that has two columns labeled time and population_total. The first column should contain the years from 1970 through 2015 (including both 1970 and 2015) and the second should contain the population of Bangladesh in each of those years.

Question 2. Create a table called b_five_growth that includes three columns, time, population_total, and annual_growth. There should be one row for every five years from 1970 through 2010 (but not 2015). The first two columns are the same as b_five. The third column is the **annual** growth rate for each five-year period. For example, the annual growth rate for 1975 is the yearly exponential growth rate that describes the total growth from 1975 to 1980 when applied 5 times.

Question 3. Perhaps population is growing more slowly because people aren't living as long. Use the life_expectancy table to draw a line graph with the years 1970 and later on the horizontal axis that shows how the *life expectancy at birth* has changed in Bangladesh.

Question 4. Does the graph above help directly explain why the population growth rate decreased from 1985 to 2010 in Bangladesh? Why or why not? What happened in Bangladesh in 1991, and does that event explain the change in population growth rate?

Question 5. Write a function fertility_over_time that takes the Alpha-3 code of a country and a start year. It returns a two-column table with labels "Year" and "Children per woman" that can be used to generate a line chart of the country's fertility rate each year, starting at the start year. The plot should include the start year and all later years that appear in the fertility table.
Then, in the next cell, call your fertility_over_time function on the Alpha-3 code for Bangladesh and the year 1970 in order to plot how Bangladesh's fertility rate has changed since 1970.

Question 6. Does the graph above help directly explain why the population growth rate decreased from 1985 to 2010 in Bangladesh? Why or why not?

Question 7. Using both the fertility and child_mortality tables, draw a scatter diagram with one point for each year, starting with 1970, that has Bangladesh's total fertility on the horizontal axis and its child mortality on the vertical axis.

Question 8. In one or two sentences, describe the association (if any) that is illustrated by this scatter diagram. Does the diagram show that reduced child mortality causes parents to choose to have fewer children?

2. Classification

You will build a classifier that guesses whether a song is hip-hop or country, using only the numbers of times words appear in the song's lyrics. By the end of the project, you should know how to:

1. Build a k-nearest-neighbors classifier.
2. Test a classifier on data.

Using code perform the following operations:

1. Load the dataset into a python Table
2. Perform word stemming on the dataset
3. Split the dataset into training and testing data.
4. Using the training data, classify a song using k-nearest-neighbors on two features
5. Extend your classifier by including 20 features
6. Write a single function that encapsulates the whole process of classification

D. OTHER TYPICAL ASSESSMENT AND EVALUATION METHODS

Objective Exam
Completion
Homework Problems
Term or Other Papers
Written Homework
Other (specify): Programming Assignments

V. INSTRUCTIONAL METHODS: Select from this list.

Lecture
Lab

Note: In compliance with Board Policies 1600 and 3410, Title 5 California Code of Regulations, the Rehabilitation Act of 1973, and Sections 504 and 508 of the Americans with Disabilities Act, instructional delivery shall provide access, full inclusion, and effective communication for students with disabilities.

VI. WORK OUTSIDE OF CLASS: Select from this list. Use all that apply.

Study
Required reading
Problem solving activity
Written work (such as essay/composition/report/analysis/research)

Estimated Study Hours Per Week: 6 hours

VII. TEXTS AND MATERIALS

A. UP-TO-DATE REPRESENTATIVE TEXTBOOKS

An Introduction to Data Science, Jeffrey Saltz; Jeffrey Stanton, SAGE Publications, Inc; First edition 2017
Computational and Inferential Thinking, Ani Adhikari; John DeNero, Retrieved October 9, 2019, from www.inferentialthinking.com.

B. ALTERNATIVE TEXTBOOKS

C. REQUIRED SUPPLEMENTARY READINGS

D. OTHER REQUIRED MATERIALS

VIII. CONDITIONS OF ENROLLMENT

A. Requisite/s (Course and Non-Course Prerequisite/s and Corequisite/s). Add rows as needed.

Requisites	Category and Justification
Course Prerequisite Mathematics-73 or	Computational/Communication Skills
Course Prerequisite Mathematics-80 or	Computational/Communication Skills
Course Prerequisite Mathematics-67 or	Computational/Communication Skills
Placement	Computational/Communication Skills

B. Requisite Skills - Match skills from prerequisite course/s or non-course prerequisites without which a student would be “highly unlikely to succeed.”

Requisite Skills – Matching
Requisite Skill Needed: Students need algebraic logic, equation, function and graphing skills, including linear and non-linear functions; numeric conversions and quantitative understanding of percents, proportions and decimals; area; expression simplifying; solving equations; order of operations, mathematical reasoning.

C. Recommended Preparations (Course and Non-Course)

Recommended Preparation	Category and Justification

D. Recommended Skills. Match skills from recommended courses or non-course prerequisite that would “enhance a students’ ability to succeed in the courses”.

Recommended Skills – Matching
(A) Recommended Skill Needed: Application Problems Students will be able to recognize and apply appropriate mathematical concepts and models involving a variety of functions to contextualized problems involving authentic, real-world data. Course title and number and objective related to that skill: MATH 67 General Education Algebra: 9. Translate problems from a variety of contexts into a mathematical representation (symbolic, tabular, and graphic) and vice versa, 7. Calculate measures of center, measures of dispersion, and measures of relative position and distinguish when to apply them appropriately. Math 73 Intermediate Algebra for General Education – 2. Recognize functional relationships in the form of graphs, data or symbolic equations. 6. Using numerical, symbolic and graphical methods, model application problems, solve them and interpret the results in the context of the problem. Math 80 Intermediate Algebra for Science, Technology, Engineering and Mathematics - 2. Recognize functional relationships in the form of graphs, data or symbolic equations. 6. Using numerical, symbolic and graphical methods, model application problems, solve them and interpret the results in the context of the problem.
(B) Recommended Skill Needed: Solving Equations and Manipulating Expressions, understanding order of operations. Math 67 - 10. Instruct and use equations and inequalities to represent relationships involving one or more unknown or variable quantities to solve problems. 2. Solve problems involving ratios, proportions and percents. Math 73 -3. Solve problems involving linear, quadratic, exponential, square root and aboslute value functions. 5. Solve a variety of equations and inequalities, as well as systems of equations and inequalities, using algebraic and graphical methods. Math 80 – 1. Carry out numerical operations and manipulate algebraic expressions, including expressions with rational and negative exponents, complex numbers, and logarithms. 3. Solve problems involving a variety of function types, including linear, quadratic, polynomial, rational, radical, exponential and logarithmic functions.

5. Solve a variety of equations and inequalities, as well as systems of equations and inequalities, using algebraic and graphical methods. Types of equations include linear, quadratic, polynomial, rational, radical, exponential and logarithmic equations.

(C) Recommended Skill Needed: Visual and Graphical Methods to represent, analyze and solve contextualized problems
Course title and number and objective related to that skill:
Math 67 – 3. Analyze simple data sets by using appropriate exploratory data analysis techniquex,
5. Construct and analyze various graphs, including bar graphs, pie charts, histograms, stem-and-leaf plots, boxplots and scatterplots.
6. Analyze reading that include quantitative or statistical information.
Math 73 – 4. Graph funcitons and use graphs to solve problems.
Math 80 – Graph a variety of funcitons and relations and draw connections between these graphs and solutions to problems.

(D) Recommended Skill Needed: Articulating Mathematical Reasoning Students should be able to think through and work logical step-by-step processes. Course title and number and objective related to that skill:
Math 67 - Recognize proportional relationships from verbal and numeric representations and compare proportional relationships represented in different ways.
2. Solve problems involving ratios, proportions and percents.
4. Calculate quantities using summation notation.
8. Present statistical results orally and in written form
after analyzing data or solving applied problems.
Math 73 - 1. Carry out numerical operations and manipulate algebraic expressoins.
Math 80 - 1. Carry out numerical operations and manipulate algebraic expressions, including expressions with rational and negative exponents, complex numbers, and logarithms.

E. Enrollment Limitations

Enrollment Limitations and Category	Enrollment Limitations Impact

Course created by: Solomon Russell, Alice Martinez Fall 2019

BOARD APPROVAL DATE: 1/21/2020

LAST BOARD APPROVAL DATE: