|  |  |
| --- | --- |
|  | Concept Assignment 8  PLTW Computer Science CSP Core Training |

# Data Visualization and Impact

|  |  |
| --- | --- |
|  | Learning Objectives |

LO8.1 While working through Activity 3.1.1: Time Series and Trends,the teacher will:

* Use *Python*® code to visualize nearly a century of baby names from all 50 states.
* Modify *Python* code to generalize data visualizations and produce custom reports.
* Use Microsoft® Excel® to calculate relative frequencies of baby name data.

LO8.2 While working through Activity 3.1.4 Pie Charts and Bar Graphs, the teacher will:

* Distinguish between data that can be best visualized using pie graphs versus bar graphs, line graphs, or box plots.
* Create pie graph visualizations with *Python* code using online documentation as a guide.
* Create side-by-side bar graphs with *Python* code*.*
* *Create side-by-side* bar graphs in an Excel spreadsheet.

LO8.3 While working through Activity 3.1.5 Histograms and Distributions, the teacher will:

* Distinguish between categorical and quantitative variables.
* Describe the axial components of a histogram.
* Use *Python* code to generate histograms.
* Understand and perform data transformations using Python code and Excel spreadsheets.
* Identify variable roles within *Python* code.
* Use *Python* code to generate box plots.
* Compare random distributions in terms of symmetry, skewing, and spread.
* Use *Python* to generate uniform distributions.

LO8.4 While working through Activity 3.2.1 Inferential Statistics, the teacher will:

* Use p-values to determine the validity of claims related to data sets.
* Make observations about data sets resulting from a pair of binary questions.
* Compare quantitative data using a visualization created with *Python* code.
* Apply concepts of linear correlation using visualizations of data sets using *Python* code.
* Explain the Law of Large Numbers.

LO8.5 While working through Activity 3.2.4 Geographic Data, the teacher will:

* Create a custom visualization of archived earthquake data using *Python* code.
* Examine various data visualizations available online, drawing conclusions and generating hypotheses regarding the underlying data.
* Discuss how data visualization impacts decision-making.

LO8.6 While working through Activity 3.2.6 Genomic Data, the teacher will

* Query over 160 million DNA sequences for matches to a selected *Homo sapien* protein.
* Using various DNA alignment tools, match a particular human protein DNA sequence with related organisms to determine the degree of relatedness between organisms.
* Generate phylogenetic trees to visualize relatedness between species.

|  |  |
| --- | --- |
|  | AP CSP Enduring Understandings (EU) and Learning Objectives (LO) |

* New knowledge can be created and communicated by using mathematics, code, creativity, and visualization of data to combine human and computer powers. EU3.1 through LO3.1.1 [P4], LO3.1.2 [P6], and LO3.1.3 [P5]; EU3.2 through LO3.2.1 [P1] and LO3.2.2 [P3]; and EU5.5 through LO5.5.1 [P1].
* Operations on data and artifacts such as visualizations can be analyzed critically. EU1.2 through LO1.2.5 [P2] and EU3.3 through LO3.3.1 [P4].
* Societal issues arise from the computational explosion of data. EU4.1 through LO4.1.1 [P2], EU5.2 through LO5.2.1 [P3].
* Some patterns are random coincidence. EU3.1 through LO3.1.1 [P4] and LO3.1.3 [P5].
* Computing is bringing to life the stuff of science fiction. EU7.1 through LO7.1.1 [P4]
* Computing is revolutionizing many fields, especially through large-scale collaboration. EU7.1 through LO7.1.2 [P4] and EU7.2 through LO7.2.1 [P1]
* Teams create and communicate new knowledge using mathematics, code, creativity, and visualization of data to combine human and computer powers. EU1.2 through LO1.2.4 [P6]; EU3.1 through LO3.1.2 [P6]; and EU3.2 through LO3.2.1 [P1] and LO3.2.2 [P3]; and EU5.1 through LO5.1.3 [P6].

|  |  |
| --- | --- |
|  | Introduction |

Computers make it possible to collect, transmit and store massive amounts of data in fractions of a second. How do we make use of all this data? And if we’re able to make sense of all that data, what kind of impact will that information have on us as individuals and as a society? These are two fundamental questions that we’ll address. For this reason, we named this concept assignment *Data Visualization and Impact*.

Data visualization is an abstraction. It’s a way we humans use our relatively superior visual abilities to make sense of numerical information. It is simply easier on our brains to detect nuance in color, shape, size, and density in data represented visually than it is to detect those same differences in a long list of numbers. Computers can help us represent massive amounts of data visually, thereby making the information behind it more accessible. Data represented visually can help us quickly recognize patterns and detect anomalies. Data visualization can also help us communicate our ideas to others more effectively. For these reasons and others, we will investigate computational tools that allow us to understand the data we’ve collected, and to appreciate the implications and consequences of doing so.

### Overview of Data Visualization and Analysis in the Course

Data is now easier to produce, explore, and analyze in large quantities because computers automate the process.Whenever your students are curious about anything, you want them to think, as a matter of habit, “How could I use computing to collect or retrieve, visualize, and analyze a large quantity or high quality of data?”

Computational thinking is a set of habits of thought about how computers can help solve a problem.

**Data skills** are one of the central sets of computational thinking skills. Data skills include being able to:

* Identify ways to collect data by automated logging or by causing data to emanate from large crowds of people.
* Identifying data that will be useful to collect and organizing the data before collection.
* Cleaning data that has been collected by correcting or removing anomalous artifacts.
* Filtering data based on Boolean conditions.
* Transforming data to produce derivative data using calculations and algorithms.
* Transforming data to be more suitable for processing, parallelism, storage, transmission, or privacy by applying encryption, changing data representation, or using compression.
* Exploring data by creating **visualizations**.
* Seeing patterns in data, asking questions, and formulating hypotheses.
* Analyzing data to infer or abstract new knowledge.

|  |  |
| --- | --- |
|  | Part 1: Review the Student Learning Sequence for Lesson 3.1 |

#### Activity 3.1.1 Time Series and Trends (3 days)

*Students will*…

Explore the frequencies of people’s names while using *Python* to read data files and visualize data. Use absolute and relative cell references in Excel to transform data. In data about the world, explore five variables at a time with animated, colorized bubble plots.

#### Activity 3.1.2 Privacy Issues with Data (2 days)

*Students will*… Know how commercial entities gather personal data, interpret privacy policies and terms of service, describe targeted advertising, and name rights of the consumer. Describe machine learning, heuristics, and recommender systems. Describe how sensitive, personally identifiable data can be de-identified and re-identified.

#### Activity 3.1.3 Data Innovations and Parallel Algorithms (2 days)

*Students will*… Use metric prefixes to describe very large data sets and describe the historical explosion of digital data. Describe a data-parallel algorithm. Describe parallel processing in computers and human brains. Research and write about the impact of innovations in data collection and computing power.

#### Activity 3.1.4 Pie Charts and Bar Graphs (3 days)

*Students will*… Create pie charts and stacked or side-by-side bar charts with *Python*. Explain how to identify the appropriate chart type for a particular set of data.

#### Activity 3.1.5 Histograms and Distributions (4 days)

*Students will*… Identify continuous and discrete variables. Create histograms to explore U.S. age and income structure using *Python*. Compare the center, spread, and skew of these distributions to the normal and uniform distributions. Become acquainted with various *Python* approaches to transform data: dictionaries, array operators, map(), and generators. Compare *Python* and Microsoft® Excel® as tools for working with data. Describe how data from a sample or a model simulation relate to population statistics.

|  |  |
| --- | --- |
|  | Part 2: Using Python for Data Visualizations |

1. Work through Activity 3.1.1 Time Series and Trends.

|  |
| --- |
| Submission Items |
| 1. Acivity 3.1.1: Produce a screenshot of the graphs from step 11 and code from step 13. 2. Activity 3.1.1: Place the artifacts from step 23 into a slideshow. |

1. Work through Activity 3.1.4 Pie Charts and Bar Graphs.

|  |
| --- |
| Submission Item |
| 1. Activity 3.1.4: Submit *Python* code and screenshots for step 3b. |

1. Work through Activity 3.1.5 Histograms and Distributions.

|  |
| --- |
| Submission Item |
| 1. Activity 3.1.5: Submit *Python* code and artifacts and screenshots for step 19. |

### Overview of Impact in the Course

Many activities, projects, and problems in CSP help students understand the impact of computing on society. Several deal specifically with social issues around data. For example, in Activity 2.3.3 Security and Liberty and Activity 3.1.3 Data Innovations and Parallel Algorithms, students identify relevant and recent information on the Web, read about innovations and societal issues, and write about the beneficial and harmful potential of computing innovations. These activities make good practice opportunities for the AP CS Principles Explore Performance Task. In this concept assignment, you’ll peruse the resources of this type that are included in the curriculum.

In other CSP activities, students explore specific innovations by using them in a hands-on manner. For example, in Activity 3.2.3 Image Data, students experiment with a *Python* application for face recognition. In this concept assignment, you’ll learn to implement Project 3.2.6 Genomic Data. This project advances student understanding, skills, and knowledge in biology as well as computer science. Like most fields, biology has been transformed by computing, and this project exposes students to the new advancements and innovative uses of computing in biology.

|  |  |
| --- | --- |
|  | Part 3: Review the Student Learning Sequence for Lesson 3.2 |

### Lesson 3.2 Student Learning Sequence

#### Activity 3.2.1 Inferential Statistics (2 days)

*Students will*…

* Explore data about teen texting and teen heights, survey their classmates, and scrape data from web pages on weather and sports.
* Create models and use Monte Carlo simulation to infer with confidence whether proportions or means in two populations are different.
* Identify linear correlations with a scatter plot array.

#### Activity 3.2.2 Linked Data (1 day)

*Students will*…

* Use terminology of graph theory to describe linked data.
* Describe breadth-first and depth-first search.
* Create poetry with the Visual Thesaurus.

#### Activity 3.2.3 Geographic Data (2 days)

*Students will*…

* Describe ways to visualize geographic data.
* Use the *Python* basemap library to create maps overlaid with data.
* Analyze geographic data to investigate patterns in baby naming, word use and pronunciation, and racial segregation.

#### Activity 3.2.4 Considering Gattaca (1 day)

*Students will*… Watch the film Gattaca at home and discuss social issues raised by government and commercial access to personal data.

#### Project 3.2.5 Genomic Data (3 days)

*Students will*…

* Describe the relationship among DNA, RNA, and protein.
* Retrieve a human DNA sequence for a protein of their choice from the National Center for Biotechnical Information (NCBI).
* Use NCBI supercomputers to find similar DNA from other species.
* Explain the algorithmic problem of aligning similar strings.
* Use the Molecular Evolution Genetics Analysis tool to analyze the similarities among the DNA data.
* Use a *Python* simulation to understand different types of genetic mutation.

#### Problem 3.2.6 Investigating with Data (6 days)

*Students will*… Find or collect data, explore and analyze the data, and communicate discoveries on a student-selected topic.

|  |  |
| --- | --- |
|  | Part 4: Impact |

1. Work through Activity 3.2.1 Inferential Statistics. Work through steps 10–14 and 18–19.

|  |
| --- |
| Submission Items |
| 1. Activity 3.2.1: Answers to questions 10–14. 2. Activity 3.2.1: Answers to question 18–19. |

1. Work through Activity 3.2.4 Geographic Data.

|  |
| --- |
| Submission Item |
| 1. Activity 3.2.4: Submit screenshots for step 11 and step 13. |

1. Work through Activity 3.2.6 Genomic Data.

|  |
| --- |
| Submission Item |
| 1. Activity 3.2.6: Submit screenshot of phylogenetic tree and Conclusion Question 7. |

|  |  |
| --- | --- |
|  | Part 5: Consider Classroom Implications |

Interdisciplinary connections within a course are known to have an especially high impact on the motivation and achievement of students who are from groups that are underrepresented in computing. Your colleagues in other disciplines might:

* Benefit from knowing when and what you’ll be teaching so that they can make the connections in their classes.
* Be willing to help you increase your expertise and comfort with the material.
* Be open to team teaching for a day or creating joint assignments or class collaborations.

Consider the following potential interdisciplinary connections:

* Project 2.2.4 – The art database could be a point of collaboration with an Art teacher.
* Project 1.1.7 – Storytelling and poetry with the Visual Thesaurus in Activity 3.2.2 and the writing components of the College Board Performance Tasks could all be points of collaboration with a Language Arts teacher. The foreign language features of the Visual Thesaurus could also point to a collaboration with a foreign language teacher.
* Activity 3.2.4 – Geologic data could be used in a partnership with an Earth Science teacher.
* Project 3.2.6 and Unit 4 – The genomics data in Project 3.2.6 and simulations available in Unit 4’s NetLogo could be points of collaboration with a Science teacher.
* Activities 3.1.1 and 3.2.4 – Data from GapMinder in Activity 3.1.1 and the use of GIS software in Activity 3.2.4 could be a point of collaboration with a Social Studies teacher.
* Activities 3.1.5 and 3.2.1 – Statistics concepts and simulations could be a point of collaboration with a Mathematics teacher.
* Project 3.2.6 – While no explicit mention of evolution is made, the phylogenetic trees can be interpreted as visualizing similarity among organisms’ DNA and they are used by evolutionary biologists to infer common ancestry. This could point to a collaboration with a Biology teacher.

1. From the list above, determine the most viable alternative at your school. Compose an email to propose such a collaboration to a colleague, using persuasive prose that spells out the benefits for that colleague and their students.
2. Speculate as to why interdisciplinary connections motivate students and speculate as to why the effect size tends to be larger for students from underrepresented groups. Reflect on your recruitment plan. How will you use the interdisciplinary nature of this course to recruit and retain diverse students?
3. Speculate as to what impact worldwide collaboration and the exchange of data (as demonstrated in Activity 3.2.4) have had on the quality and growth of biological knowledge in the world since the advent of the personal computer. Find data to support your assertions. Using a search engine of your choice, discover, cite, and share your discoveries of additional fields of study that have been similarly affected through information crowdsourcing or data sharing.

|  |
| --- |
| Submission Item |
| 1. Write a reflection about the things you learned today. Consider highlighting new things you learned, items you need to consider for implementing this in your classroom, and ideas and suggestions you heard from others. Use the questions in Part 5 of the assignment as prompts, but don’t feel limited or constrained by just those questions. |

**Trademark Attribution**  PLTW, Project Lead The Way, and the PLTW logo are registered trademarks of Project Lead The Way, Inc. “Python” and the Python logos are trademarks or registered trademarks of the Python Software Foundation. Microsoft and Excel are registered trademarks of Microsoft Corporation in the United States and/or other countries. All other brand names, product names, or trademarks belong to their respective holders.