SAN DIEGO COMMUNITY COLLEGE DISTRICT
CONTINUING EDUCATION
COURSE OUTLINE

SECTION I

SUBJECT AREA AND COURSE NUMBER

COMP 663

COURSE TITLE

PYTHON FOR DATA SCIENCE

TYPE COURSE

NON-FEE VOCATIONAL

CATALOG COURSE DESCRIPTION

This course explores the theory and concepts of data science while acquiring Python programming knowledge to solve real world data challenges. At the end of this program, students will make sense of the data by using Python's wide variety of data analytics and graphical modeling packages to perform exploratory data analysis, apply visualization and inferential techniques, as well as data mining algorithms, to real-world data that is engaging and relevant in the industry in the years ahead. (FT)

LECTURE/LABORATORY HOURS

126

ADVISORIES

COMP 660 PROGRAMMING WITH PYTHON I; and
COMP 661 PROGRAMMING WITH PYTHON II; and
COMP 662 PYTHON FOR DATA MANAGEMENT

RECOMMENDED SKILL LEVEL

- Possess a 12th grade reading level
- Ability to communicate effectively in the English language
- Knowledge of math concepts at the 8th grade level and computer literacy

INSTITUTIONAL STUDENT LEARNING OUTCOMES

1. Social Responsibility
   SDCE students demonstrate interpersonal skills by learning and working cooperatively in a diverse environment.
2. Effective Communication
   SDCE students demonstrate effective communication skills.

INSTITUTIONAL STUDENT LEARNING OUTCOMES (CONTINUED)

3. Critical Thinking
   SDCE students critically process information, make decisions, and solve problems independently or cooperatively.
4. Personal and Professional Development
   SDCE students pursue short term and life-long learning goals, mastering necessary skills and using resource management and self-advocacy skills to cope with changing situations in their lives.

COURSE GOALS

1. Learn critical concepts required to enter the world of Data Science via Python and develop relevant programming abilities
2. Demonstrate proficiency with statistical analysis of data
3. Develop the ability to build and evaluate data-based models
4. Execute statistical analyses with some of the most widely used Python packages
5. Demonstrate advanced skill in data management, integrate data from disparate sources, transform data from one format to another
6. Work with big data technologies to manage varied data at high velocity and volume
7. Create high-end visualizations using Python
8. Build recommendation engine models with various collaborative filtering algorithms
9. Apply data science concepts and techniques to solve problems in real-world circumstances and communicate these solutions effectively

COURSE OBJECTIVES

Upon successful completion of the course, the student will be able to:

1. Explain what data science is, the diverse activities of a data scientist's job, and the methodology to analyze and work as a data scientist
2. Demonstrate hands-on skills using the tools, languages, and libraries used by professional data scientists
3. Import structured and unstructured data into Python and parse unstructured data into structured formats by data cleansing and preparation.
4. Understand mechanisms for missing data and handle by removing records and imputation, feature scaling using standardization and normalization and dimensionality reduction
5. Use iPython notebooks for ad hoc calculations, plots, and what-if analysis
6. Explain the fundamentals of some of the most widely used Python packages; including NumPy, pandas, and Matplotlib, then apply them to solve equations for data analysis and visualizations
7. Import, clean, enrich, transform, visualize, and output the analysis of a large dataset with different types of bar charts, pie charts, scatter plots, and line charts
8. Display distribution of data with box plots, histograms, and violin plots
9. Import and clean data sets, analyze and visualize data, build and evaluate machine learning models and pipelines using Python
10. Implement supervised machine learning categorization for classification and regression models and implement unsupervised machine learning categorization for clustering and anomaly detection

SECTION II

COURSE CONTENT AND SCOPE

1. Applied Data Science
   1.1. Data science defined
   1.2. The data science ecosystem
   1.3. Data mining vs. data science
   1.4. Business analytics vs. data science
   1.5. Data science, machine learning, and artificial intelligence (AI)
   1.6. Defining the role of a data scientist
   1.7. Data scientists at work
2. Python For Data Science
   2.1. The Python data science “Ecosystem”
   2.2. NumPy
   2.3. NumPy arrays
   2.4. NumPy idioms
   2.5. pandas
   2.6. Data wrangling with pandas DataFrame
   2.7. SciPy
   2.8. Scikit-learn
   2.9. Matplotlib
   2.10. Python vs R
   2.11. Anaconda
   2.12. IPython
   2.13. Visual studio code
   2.14. Jupyter
3. Data Analytics Life-Cycle
   3.1. Big data analytics pipeline
   3.2. Data discovery phase
   3.3. Data harvesting phase
   3.4. Data priming phase
   3.5. Data logistics and data governance
   3.6. Exploratory data analysis
   3.7. Model planning phase
   3.8. Model building phase
4. Repairing And Normalizing Data
   4.1. Repairing and normalizing data
   4.2. Dealing with the missing data
   4.3. Sample data set
   4.4. Getting info on null data
   4.5. Dropping a column
   4.6. Interpolating missing data in pandas
   4.7. Replacing the missing values with the mean value
   4.8. Scaling (Normalizing) the data
   4.9. Filtering data with pandas query()
   4.10. Evaluate Python expressions with pandas eval()

COURSE CONTENT AND SCOPE (CONTINUED)

5. Descriptive Statistics Computing Features In Python
   5.1. Descriptive statistics
   5.2. Non-uniformity of a probability distribution
   5.3. Finding min and max in NumPy
   5.4. Using pandas for calculating descriptive statistics measures
   5.5. Regression and correlation
   5.6. Finding min and max in pandas DataFrame
6. Data Aggregation And Grouping
   6.1. Working with pandas
   6.2. Data aggregation and grouping
   6.3. Grouping by two or more columns
   6.4. Emulating the structured Query Language's WHERE clause
   6.5. The pivot tables
   6.6. Cross-tabulation
7. Data Visualization With Matplotlib
   7.1. Working in Jupyter notebooks
   7.2. Data visualization
   7.3. The plotting window
   7.4. The figure options
   7.5. Customizing plot legends
   7.6. Customizing ticks
   7.7. Subplots
   7.8. Histograms, binnings, and density
   7.9. Text and annotation
   7.10. The matplotlib.pyplot.subplot() function
   7.11. Saving figures to file
   7.12. Visualization with Matplotlib
   7.13. Visualization with Seaborn
8. Machine Learning
   8.1. Data science, machine learning, AI?
   8.2. Types of machine learning
   8.3. The scikit-learn package
   8.4. Models, estimators, and predictors
   8.5. Supervised machine learning algorithms
   8.6. Unsupervised machine learning algorithms
   8.7. Data split for training and test data sets
   8.8. Decision tree classification in context of information theory
   8.9. Bayes formula
   8.10. Time-Series analysis

APPROPRIATE READINGS

Reading assignments may include, but are not limited to assigned readings from textbooks, supplemental reading assignments, industry-related periodicals or magazines, manuals, online help pages, articles posted on the Internet, and information from web sites, online libraries and databases. Topics should be related to Python programming with data science to start or continue your data science journey to include techniques for repairing and normalizing, aggregation and grouping or visualization of data using Python.

WRITING ASSIGNMENTS

Writing assignments may include, but are not limited to, completing assigned reports, providing written answers to assigned questions, performing internet research and reporting on that research. An example would include a case study about how an organization uses Python and data science to run basic inferential statistical analysis and communicate insights covering big data framework and applications

OUTSIDE ASSIGNMENTS

Outside assignments may include, but are not limited to, appropriate internet research, reading from assigned textbooks and completing the assignments at the end of each chapter, and studying as needed to perform successfully in class. An appropriate assignment for instance, would include the creation of an application that stores support requests made from clients for an organization’s customer support department.

APPROPRIATE ASSIGNMENTS THAT DEMONSTRATE CRITICAL THINKING

Assignments, which demonstrate critical thinking, may include but are not limited to designing and building an application with applied focus on Python for data science using interactive open-source platforms for computational analysis from a wide variety of industries such as manufacturing, retail, financial services, ecommerce, financial technology, and healthcare. Students will also be expected to participate in online class discussion posts, in-class discussions, and project reviews.

EVALUATION

Evaluation that a student has met the course competencies will include multiple measures of performance related to the course objectives. Evaluation methods may include, but are not limited to performance in a variety of activities and assignments, such as completing a research project individually or in a group, hands-on projects, and demonstration of use of the internet, quizzes, class participation, written and practical tests, attendance and punctuality.

Upon successful completion of all courses in the program, a Certificate of Program Completion will be issued.

METHOD OF INSTRUCTION

Methods of instruction may include, but are not limited to, lecture, in-class and online discussions, hands-on demonstrations, computer-assisted instruction, field trips, and laboratory assignments.

This course, or sections of this course, may be offered through distance education.

TEXTS AND SUPPLIES

Open Educational Resources (OER) Textbooks:

- *Python Data Science Handbook, V*anderPlas, J.). O'Reilly, 2016
- *Automate the Boring Stuff with Python,* Al Sweigart, No Starch Press, current edition
- *Think Stats 2e., A*llen Downey Green Tea Press, 2015 (http://greenteapress.com/thinkstats2/thinkstats2.pdf)
- *Executive Data Science A Guide to Training and Managing the Best Data Scientists,* Brian Caffo, Roger D. Peng and Jeffrey Leek. Lean Pub (https://leanpub.com/eds) , 2018-12-12

PREPARED BY Alexander Wassell DATE January 6, 2021

REVISED BY _____ DATE _____

Instructors must meet all requirements stated in Policy 3100 (Student Rights, Responsibilities and Administrative Due Process), and the Attendance Policy set forth in the Continuing Education Catalog.

REFERENCES:

San Diego Community College District Policy 3100
California Community Colleges, Title 5, Section 55002
Continuing Education Catalog