Python for Data Analysis 3rd edition McKinney Test Bank

SKU: 9781098104030-TEST-BANK

Python for Data Analysis

ISBN 9781098104030 | Chapter 1: Preliminaries

Total Questions: 150

Multiple Choice

Q1.

What is the primary focus of the book described in the chapter?

A) Advanced statistical theory for experimental design

B) Python programming, libraries, and tools for data analysis ✓ Correct

C) Database administration for enterprise systems

D) Machine learning model deployment only

Rationale: The chapter states that the book focuses specifically on Python programming, libraries, and tools used for data analysis.

Q2.

In the chapter, the terms wrangling and munging most closely refer to which activity?

A) Data manipulation ✓ Correct

B) Model validation

C) Network configuration

D) Compiler optimization

Rationale: The text explicitly notes that wrangling or munging are terms used for data manipulation.

Q3.

Which type of data is identified as a primary focus of the book?

A) Unstructured audio streams only

B) Structured data ✓ Correct

C) Genomic sequence data only

D) Binary executable data

Rationale: The chapter says the primary focus is on structured data in various common forms.

Q4.

Which of the following is listed as an example of structured data?

A) A handwritten note image with no extracted features

B) A raw video file with no metadata

C) A tabular dataset with columns of different types ✓ Correct

D) A compiled operating system kernel

Rationale: Tabular or spreadsheet-like data with mixed column types is given as a structured data example.

Python for Data Analysis 3rd edition McKinney Test Bank

Q5.

Which statement best describes Python according to the chapter?

A) It is only suitable for writing short automation scripts
B) It is a compiled language primarily used for firmware
C) It is an interpreted language used for both scripting and serious software development ✓ Correct
D) It is limited to academic use and not industry

Rationale: The author describes Python as an interpreted language and rejects the idea that it cannot be used for serious software.

Q6.

Python first appeared in which year according to the chapter?

A) 1985
B) 1991 ✓ Correct
C) 2001
D) 2010

Rationale: The chapter states that Python first appeared in 1991.

Q7.

What is meant by Python being used as glue?

A) It is mainly used to create graphical interfaces
B) It is used to replace all low-level languages in scientific computing
C) It is used to integrate components such as legacy C, C++, and FORTRAN code ✓ Correct
D) It is used only for packaging software distributions

Rationale: Python as glue refers to its ease of integrating and coordinating existing lower-level codebases.

Q8.

In the context of the chapter, what is the two-language problem?

A) The difficulty of translating Python into SQL automatically
B) The need to use one language for research and another for production systems ✓ Correct
C) The challenge of writing multilingual documentation
D) The inability of Python to interface with C libraries

Rationale: The chapter defines the two-language problem as prototyping in one language and later porting to another for production.

Q9.

Which technology is mentioned as helping achieve excellent performance in computational algorithms without leaving Python?

A) Apache Ant
B) Numba JIT compilation ✓ Correct
C) SOAP web services
D) HTML templating

Rationale: The chapter cites just-in-time compiler technology from libraries like Numba.

Python for Data Analysis 3rd edition McKinney Test Bank

Q10.

Which feature is identified as a challenge for highly concurrent, CPU-bound multithreaded Python applications?

A) The global interpreter lock ✓ Correct
B) Automatic memory mapping
C) Columnar storage
D) Vectorized broadcasting

Rationale: The chapter explains that the GIL prevents execution of more than one Python instruction at a time.

Q11.

What is NumPy primarily known for in the Python data ecosystem?

A) Web application routing
B) Numerical computing with fast multidimensional arrays ✓ Correct
C) Version control integration
D) Authentication and authorization

Rationale: NumPy is described as a cornerstone of numerical computing centered on the ndarray.

Q12.

What is the name of NumPy's multidimensional array object?

A) DataFrame
B) Series
C) ndarray ✓ Correct
D) Panel

Rationale: The chapter explicitly names ndarray as NumPy's multidimensional array object.

Q13.

Which library is described as providing high-level data structures for structured or tabular data?

A) pandas ✓ Correct
B) matplotlib
C) IPython
D) Patsy

Rationale: pandas is introduced as providing high-level structures and functions for structured or tabular data.

Q14.

Which pair of primary pandas objects is emphasized in the chapter?

A) Array and Matrix
B) Table and Vector
C) DataFrame and Series ✓ Correct
D) Record and Field

Rationale: The chapter identifies DataFrame and Series as the primary pandas objects used in the book.

Python for Data Analysis 3rd edition McKinney Test Bank

Q15.

What is a key benefit of pandas data structures with labeled axes?

A) They eliminate the need for file storage
B) They prevent common errors from misaligned data ✓ Correct
C) They compile Python code to machine code
D) They replace relational databases entirely

Rationale: The author notes that labeled axes and alignment help prevent errors due to misaligned data.

Q16.

The name pandas is derived in part from which term?

A) Parallel analytics data systems
B) Panel data ✓ Correct
C) Pattern and distribution sampling
D) Python array decomposition structures

Rationale: The chapter states that pandas is derived from panel data and the phrase Python data analysis.

Q17.

Which library is identified as the most popular Python library for producing plots and other two-dimensional visualizations?

A) statsmodels
B) matplotlib ✓ Correct
C) scikit-learn
D) Numba

Rationale: matplotlib is described as the most popular Python library for two-dimensional plotting.

Q18.

What was the original purpose of the IPython project?

A) To build a distributed database engine
B) To create a better interactive Python interpreter ✓ Correct
C) To provide statistical formula syntax
D) To replace the Python compiler

Rationale: The chapter says IPython began as Fernando Pérez's project to make a better interactive Python interpreter.

Q19.

Which workflow does IPython encourage according to the chapter?

A) Edit-compile-link-deploy
B) Batch-submit-review
C) Execute-explore ✓ Correct
D) Design-document-approve

Rationale: The text contrasts IPython's execute-explore workflow with the typical edit-compile-run workflow.

Python for Data Analysis 3rd edition McKinney Test Bank

Q20.

What did the IPython web notebook become after the Jupyter project was announced?

A) The SciPy console
B) The pandas studio
C) The Jupyter notebook ✓ Correct
D) The NumPy browser

Rationale: The chapter states that the IPython web notebook became the Jupyter notebook.

Q21.

Which of the following can Jupyter notebooks author in addition to code?

A) Only binary executables
B) Markdown and HTML ✓ Correct
C) Only SQL schema files
D) Only XML configuration files

Rationale: The chapter notes that Jupyter notebooks allow authoring content in Markdown and HTML.

Q22.

SciPy is best described as which of the following?

A) A package manager for Python environments
B) A collection of packages for foundational scientific computing problems ✓ Correct
C) A web framework for dashboards
D) A replacement for the Python standard library

Rationale: SciPy is introduced as a collection of packages addressing foundational scientific computing problems.

Q23.

Which SciPy module is associated with function optimizers and root-finding algorithms?

A) scipy.signal
B) scipy.optimize ✓ Correct
C) scipy.stats
D) scipy.sparse

Rationale: The chapter lists scipy.optimize as containing minimizers and root-finding algorithms.

Q24.

Which SciPy module includes standard probability distributions and statistical tests?

A) scipy.linalg
B) scipy.integrate
C) scipy.stats ✓ Correct
D) scipy.special

Rationale: scipy.stats is described as containing distributions, samplers, tests, and descriptive statistics.

Python for Data Analysis 3rd edition McKinney Test Bank

Q25.

How is scikit-learn characterized in the chapter?

A) The premier general-purpose machine learning toolkit for Python programmers ✓ Correct
B) A package exclusively for symbolic mathematics
C) A package manager for Linux distributions
D) A spreadsheet replacement for accountants only

Rationale: The chapter describes scikit-learn as the premier general-purpose machine learning toolkit for Python.

Q26.

Which of the following is listed as a scikit-learn model selection capability?

A) ANOVA tables
B) Grid search and cross-validation ✓ Correct
C) ARIMA modeling
D) Kernel density estimation

Rationale: The scikit-learn section explicitly lists grid search, cross-validation, and metrics under model selection.

Q27.

Compared with scikit-learn, statsmodels is more focused on which objective?

A) GPU acceleration
B) Statistical inference ✓ Correct
C) Web scraping
D) Interactive plotting only

Rationale: The chapter states that statsmodels emphasizes inference, including uncertainty estimates and p-values.

Q28.

Which package provides a formula or model specification framework for statsmodels inspired by R?

A) seaborn
B) Patsy ✓ Correct
C) Numba
D) openpyxl

Rationale: The chapter identifies Patsy as the formula framework used with statsmodels.

Q29.

Which installation approach does the chapter recommend for reproducible setup across operating systems?

A) System package managers only
B) Manual compilation from source for every package
C) Miniconda with conda-forge ✓ Correct
D) Using only pip in the base Python installation

Rationale: The author recommends Miniconda with the conda-forge channel for setup.

Python for Data Analysis 3rd edition McKinney Test Bank

Q30.
What Python version does the book use throughout according to the chapter?

A) 2.7
B) 3.8
C) 3.9
D) 3.10 ✓ Correct
Rationale: The installation section states that the book uses Python 3.10 throughout.

Q31.
After configuring conda-forge, which command is used to create the environment named pydata-book with Python 3.10?

A) conda install -n pydata-book python=3.10
B) conda create -y -n pydata-book python=3.10 ✓ Correct
C) pip create pydata-book python=3.10
D) python -m venv pydata-book=3.10
Rationale: The chapter gives the exact command conda create -y -n pydata-book python=3.10.

Q32.
Which command activates the newly created conda environment for the book?

A) conda start pydata-book
B) conda enable pydata-book
C) conda activate pydata-book ✓ Correct
D) python activate pydata-book
Rationale: The chapter explicitly instructs users to activate the environment with conda activate pydata-book.

Q33.
Which package installation method should generally be preferred when using Miniconda?

A) easy_install
B) Manual wheel extraction
C) conda install ✓ Correct
D) Source compilation with make
Rationale: The chapter advises that conda install should always be preferred when using Miniconda.

Q34.
If a package is unavailable through conda, what does the chapter recommend trying next?

A) apt install
B) pip install ✓ Correct
C) brew install
D) svn checkout
Rationale: The text states that if conda install fails, users should try pip install for that package.

Python for Data Analysis 3rd edition McKinney Test Bank

Q35.
What caution is given about using conda and pip in the same environment?

A) They cannot install any of the same package names
B) They should be alternated for every package update
C) Updating packages installed by one tool with the other can cause environment problems ✓ Correct
D) pip should always overwrite conda packages for compatibility

Rationale: The chapter warns against updating packages installed with one tool by using the other.

Q36.
Which environment does the author say he almost always uses as a standard development setup?

A) A browser-only notebook with no shell
B) IPython plus a text editor ✓ Correct
C) A Java IDE with Python plugins
D) A database console with spreadsheet macros

Rationale: The author explicitly says his standard development environment is IPython plus a text editor.

Q37.
Which of the following is listed as a free IDE currently shipped with Anaconda?

A) Komodo IDE
B) PyCharm Professional
C) Spyder ✓ Correct
D) MATLAB

Rationale: Spyder is identified as a free IDE currently shipped with Anaconda.

Q38.
Which mailing list or group is specifically mentioned for NumPy-related questions?

A) pydata
B) numpy-discussion ✓ Correct
C) pystatsmodels
D) scikit-image-users

Rationale: The chapter lists numpy-discussion for NumPy-related questions.

Q39.
Which conference series is described as targeted at data science and data analysis use cases?

A) EuroPython only
B) PyData ✓ Correct
C) DEF CON
D) SIGGRAPH

Rationale: The chapter describes PyData as a worldwide series of regional conferences for data science and data analysis.

Python for Data Analysis 3rd edition McKinney Test Bank

Q40.

For readers new to Python, which chapters are identified as prerequisite material?

A) Chapters 2 and 3 ✓ Correct
B) Chapters 4 and 5
C) Appendix A only
D) Chapters 10 and 11

Rationale: The chapter advises new programmers to spend time in Chapters 2 and 3 on Python features and IPython/Jupyter.

Q41.

Which broad task group in the chapter includes cleaning, munging, combining, normalizing, reshaping, and transforming data?

A) Presentation
B) Preparation ✓ Correct
C) Modeling and computation
D) Interacting with the outside world

Rationale: These activities are listed under the broad group called Preparation.

Q42.

In the code example notation shown in the chapter, what does the In block represent?

A) The package installation log
B) The code the reader is intended to execute ✓ Correct
C) The operating system shell history only
D) The final published figure output

Rationale: The chapter says the intent is for the reader to type and execute the code shown in the In block.

Q43.

Which pandas display option is set to 80 in the chapter's output-formatting example?

A) pd.options.display.max_rows
B) pd.options.display.precision
C) pd.options.display.max_colwidth ✓ Correct
D) pd.options.display.width

Rationale: The example sets pd.options.display.max_colwidth = 80.

Q44.

Which NumPy print option is set in the chapter to improve readability of output?

A) np.seterr(all='ignore')
B) np.set_printoptions(precision=4, suppress=True) ✓ Correct
C) np.show_config(precision=4)
D) np.printoptions(width=80)

Rationale: The chapter provides the exact statement np.set_printoptions(precision=4, suppress=True).

Python for Data Analysis 3rd edition McKinney Test Bank

Q45.
Where are the datasets for the chapter examples hosted according to the text?

A) Only on a private FTP server
B) In a GitHub repository, with a mirror on Gitee ✓ Correct
C) Only on the Python Package Index
D) Embedded directly in the Python interpreter

Rationale: The chapter states that example datasets are hosted on GitHub or mirrored on Gitee.

Q46.
If a reader downloads a zip file of the example datasets, what must be done before running the code examples?

A) Convert the files to CSV format
B) Import the zip directly into pandas
C) Fully extract the contents and navigate to that directory from the terminal ✓ Correct
D) Rename all notebooks to .py files

Rationale: The text instructs readers to fully extract the zip contents and navigate to that directory.

Q47.
Which import convention is presented as standard for NumPy?

A) import numpy as num
B) import np as numpy
C) import numpy as np ✓ Correct
D) from numpy import all

Rationale: The import conventions section lists import numpy as np.

Q48.
Why does the chapter discourage using from numpy import * from a large package?

A) It prevents numerical computation
B) It is considered bad practice in Python software development ✓ Correct
C) It disables array broadcasting
D) It is incompatible with Jupyter notebooks

Rationale: The chapter states that importing everything from a large package like NumPy is considered bad practice.

Q49.
A learner wants a Python tool that emphasizes prediction-focused machine learning rather than uncertainty estimates and p-values. Which library best fits that description?

A) statsmodels
B) scikit-learn ✓ Correct
C) matplotlib
D) Patsy

Rationale: The chapter contrasts scikit-learn as prediction focused, whereas statsmodels emphasizes inference.

Python for Data Analysis 3rd edition McKinney Test Bank

Q50.
A user plans to install Miniconda on a recent Apple Silicon Mac. Which installer naming pattern best matches the chapter's recommendation?

A) Miniconda3-latest-Windows-x86.exe
B) Miniconda3-latest-MacOSX-arm64.sh ✓ Correct
C) Miniconda3-latest-Linux-x86_64.sh
D) Miniconda3-latest-MacOSX-ppc64.sh

Rationale: The macOS section specifies the arm64 installer for Apple Silicon-based Macs.

Q51.
A health informatics analyst needs to clean encounter extracts from CSV files, join them to a provider lookup table by key columns, and summarize monthly counts before modeling. Which Python library should be the primary tool for this workflow?

A) matplotlib
B) pandas ✓ Correct
C) scikit-learn
D) statsmodels

Rationale: pandas is designed for structured tabular data manipulation, merging, cleaning, reshaping, and aggregation.

Q52.
A quality improvement team wants to explore patient satisfaction data interactively, trying code snippets, inspecting results immediately, and documenting findings with narrative text and code in one place. Which environment best supports this workflow?

A) Jupyter notebook ✓ Correct
B) C++ compiler
C) FORTRAN shell
D) SQL Server Agent

Rationale: Jupyter notebooks support interactive execution along with Markdown and HTML for narrative documentation.

Q53.
A data engineer at a hospital is deciding whether Python is appropriate for a new analytics service that must respond with extremely low latency in a resource-constrained environment. Based on the chapter, what is the best recommendation?

A) Use Python because interpreted code is always faster to optimize operationally
B) Use Python because the GIL improves CPU-bound throughput
C) Consider a lower-level compiled language because Python may be less suitable for very low-latency systems ✓ Correct
D) Avoid compiled languages because programmer time is always more important than CPU time

Rationale: The chapter notes that applications with very low latency or demanding resource requirements may justify using a lower-level compiled language.

Python for Data Analysis 3rd edition McKinney Test Bank

Q54.
A research department prototypes risk models in one language and then reimplements them in another for production systems, causing delays and inconsistency. Which concept from the chapter most directly addresses this problem?

A) Global interpreter lock
B) Two-language problem ✓ Correct
C) Sparse matrix decomposition
D) Panel data alignment

Rationale: The two-language problem refers to using one language for prototyping and another for production, which Python can help reduce.

Q55.
A clinical analytics team has legacy C and FORTRAN code for optimization and linear algebra but wants a Python-based workflow for orchestration and reporting. Which description best matches Python's role here?

A) Python as a plotting backend
B) Python as glue ✓ Correct
C) Python as a database engine
D) Python as a markup language

Rationale: Python as glue refers to integrating legacy lower-level code and libraries into a unified workflow.

Q56.
A nurse informaticist is combining medication administration records from two systems with different row labels and timestamps and wants tooling that reduces errors from misaligned data. Which pandas capability is most relevant?

A) Automatic or explicit data alignment with labeled axes ✓ Correct
B) Native GPU training for neural networks
C) Direct replacement of all SQL databases
D) Compiled bytecode optimization for loops

Rationale: pandas was designed with labeled axes and alignment features to prevent errors from misaligned data.

Q57.
A public health team needs one-dimensional labeled data for a single measure tracked by date, rather than a full table. Which pandas object is the most appropriate?

A) ndarray
B) Series ✓ Correct
C) DataFrame
D) Figure

Rationale: A Series is pandas' one-dimensional labeled array object.

Python for Data Analysis 3rd edition McKinney Test Bank

Q58.
A hospital analyst needs a two-dimensional labeled table with row and column labels for admissions, diagnoses, and discharge dates. Which pandas object should be used?

A) DataFrame ✓ Correct
B) Series
C) Kernel
D) Shell profile

Rationale: A DataFrame is pandas' tabular, column-oriented structure with both row and column labels.

Q59.
A biomedical researcher needs fast multidimensional numerical arrays, element-wise arithmetic, and linear algebra support as a foundation for scientific computing. Which library is the best fit?

A) NumPy ✓ Correct
B) IPython
C) matplotlib
D) beautifulsoup4

Rationale: NumPy provides efficient multidimensional arrays and core numerical computing functionality.

Q60.
A team is building a prediction model for readmission risk and needs tools for classification, preprocessing, cross-validation, and grid search. Which package is the most appropriate primary choice?

A) statsmodels
B) scikit-learn ✓ Correct
C) matplotlib
D) patsy

Rationale: scikit-learn is the general-purpose machine learning toolkit covering classification, preprocessing, and model selection.

Q61.
A doctoral student needs coefficient uncertainty estimates and p-values for a regression analyzing staffing levels and patient outcomes. Which package is most aligned with this goal?

A) scikit-learn
B) statsmodels ✓ Correct
C) matplotlib
D) Jupyter

Rationale: statsmodels emphasizes statistical inference, including uncertainty estimates and p-values.

Python for Data Analysis 3rd edition McKinney Test Bank

Q62.

A data scientist wants to create publication-quality two-dimensional visualizations of infection trends over time. Which library is the standard default choice described in the chapter?

A) matplotlib ✓ Correct

B) NumPy

C) SciPy

D) PyTorch

Rationale: matplotlib is described as the most popular Python library for two-dimensional plotting and publication-quality graphics.

Q63.

A healthcare operations analyst is using Miniconda and wants to install packages in the recommended way. Which approach should generally be preferred?

A) Use pip first for every package

B) Use conda install first and use pip only when a package is unavailable through conda ✓ Correct

C) Use apt instead of any Python package manager on all systems

D) Mix pip and conda updates freely within the same environment

Rationale: The chapter recommends preferring conda install when using Miniconda and falling back to pip only if needed.

Q64.

A student installed Miniconda on Windows but chose not to add it to the system PATH. How should the student most directly verify that the installation works?

A) Open the Windows registry editor and inspect environment keys

B) Launch the Anaconda Prompt from the Start menu and run python ✓ Correct

C) Run gcc from PowerShell

D) Open Excel and import a CSV file

Rationale: If Miniconda is not added to PATH on Windows, the chapter recommends using the Start menu Anaconda Prompt and launching python there.

Q65.

A Linux user wants a reproducible Python setup across distributions and easier access to current package versions for the book. Which installation approach best matches the chapter's recommendation?

A) Install only from each distribution's default package manager

B) Compile Python and every library manually from source

C) Use Miniconda with conda-forge ✓ Correct

D) Use a Java virtual machine with Python bindings

Rationale: The chapter recommends Miniconda with conda-forge for reproducibility across Linux distributions and simpler upgrades.

Python for Data Analysis 3rd edition McKinney Test Bank

Q66.
A macOS user with Apple Silicon hardware is selecting a Miniconda installer. Which file type is the correct match?

A) Miniconda3-latest-MacOSX-x86_64.sh
B) Miniconda3-latest-MacOSX-arm64.sh ✓ Correct
C) Miniconda3-latest-Linux-x86_64.sh
D) Miniconda3-latest-Windows-x86.exe

Rationale: Apple Silicon Macs should use the arm64 macOS installer.

Q67.
After creating a conda environment named pydata-book, a student closes the terminal and later reopens it. What must the student do before running the book's code in that environment?

A) Reinstall Python 3.10
B) Run conda activate pydata-book ✓ Correct
C) Run pip freeze
D) Edit the shell profile manually every time

Rationale: The chapter states that conda activate must be used each time a new terminal is opened.

Q68.
A student wants to follow the book using the minimum essential packages first. Which command best matches the chapter's initial package installation recommendation after activating the environment?

A) conda install -y pandas jupyter matplotlib ✓ Correct
B) pip install tensorflow pytorch seaborn
C) conda update python pandas scipy
D) apt install numpy pandas jupyterlab

Rationale: The initial essential package set listed is pandas, jupyter, and matplotlib via conda install.

Q69.
An analyst used conda to install pandas and later plans to update it. Which action is most consistent with the chapter's caution?

A) Update pandas with pip install --upgrade pandas regardless of how it was installed
B) Update pandas with conda update pandas ✓ Correct
C) Uninstall Miniconda and reinstall Python from source
D) Use both conda update and pip install --upgrade in sequence

Rationale: Packages installed with conda should generally be updated with conda to avoid environment problems.

Q70.
A team member needs to know which conda environment is active during a troubleshooting session. Which command is specifically mentioned for viewing this information?

A) conda info ✓ Correct
B) python --version
C) pip list
D) jupyter kernelspec list

Rationale: The chapter notes that conda info can be used to see information about the active environment.

Python for Data Analysis 3rd edition McKinney Test Bank

Q71.
A clinical data scientist is evaluating whether Python is ideal for a CPU-bound application using many threads in a single process. Which limitation from the chapter is most relevant?

A) Lack of any statistical libraries
B) Absence of visualization support
C) Global interpreter lock affecting CPU-bound multithreaded applications ✓ Correct
D) Inability to read files from disk

Rationale: The GIL can make highly concurrent CPU-bound multithreaded applications challenging in Python.

Q72.
A software engineer says Python can never run truly parallel code. Based on the chapter, which response is most accurate?

A) Correct, Python always executes only one instruction across all languages and extensions
B) Incorrect, Python C extensions using native multithreading can run in parallel if they do not need frequent interaction with Python objects ✓ Correct
C) Incorrect, any pure Python loop automatically bypasses the GIL
D) Correct, parallel execution is possible only in Java

Rationale: The chapter explains that C or C++ extensions can execute parallel code without being impacted by the GIL under certain conditions.

Q73.
A quantitative analyst wants better performance for computational algorithms without leaving the Python environment to write C++. Which option from the chapter most directly supports this goal?

A) Markdown
B) Numba JIT compilation ✓ Correct
C) HTML rendering
D) Patsy formulas

Rationale: Numba is cited as providing JIT compiler technology that can deliver excellent performance while staying in Python.

Q74.
A hospital finance analyst needs to fit a model and then focus primarily on prediction accuracy rather than parameter inference. Which package orientation best matches this need?

A) statsmodels, because it is more prediction focused
B) scikit-learn, because it is more prediction focused ✓ Correct
C) matplotlib, because it optimizes model selection
D) IPython, because it estimates p-values

Rationale: The chapter contrasts scikit-learn as prediction focused versus statsmodels as inference focused.

Python for Data Analysis 3rd edition McKinney Test Bank

Q75.
A biostatistics student needs to specify regression models using a formula framework inspired by R. Which project associated with statsmodels provides this capability?

A) Patsy ✓ Correct
B) Numba
C) seaborn
D) PyDev

Rationale: Patsy provides a formula or model specification framework for statsmodels inspired by R.

Q76.
A team is storing large numerical matrices and passing them among multiple lower-level scientific libraries without unnecessary copying. Why are NumPy arrays especially useful in this scenario?

A) They automatically generate p-values for each element
B) They serve as an efficient container interoperable with lower-level code ✓ Correct
C) They replace all relational operations found in SQL
D) They require no memory representation

Rationale: NumPy arrays are efficient for numerical data and can be used by C or FORTRAN libraries without copying into another representation.

Q77.
A health services researcher needs numerical integration routines and differential equation solvers for a disease progression model. Which SciPy submodule is the best match?

A) scipy.signal
B) scipy.optimize
C) scipy.integrate ✓ Correct
D) scipy.stats

Rationale: scipy.integrate contains numerical integration routines and differential equation solvers.

Q78.
An imaging analyst is working with sparse matrices from a large clinical feature set and needs sparse linear system solvers. Which SciPy submodule is most appropriate?

A) scipy.sparse ✓ Correct
B) scipy.special
C) scipy.linalg
D) scipy.stats

Rationale: scipy.sparse provides sparse matrices and sparse linear system solvers.

Q79.
A researcher needs standard probability distributions, statistical tests, and descriptive statistics for a trial dataset. Which SciPy submodule should be used?

A) scipy.optimize
B) scipy.stats ✓ Correct
C) scipy.signal
D) scipy.integrate

Rationale: scipy.stats includes probability distributions, statistical tests, and descriptive statistics.

Python for Data Analysis 3rd edition McKinney Test Bank

Q80.
A machine learning practitioner needs dimensionality reduction and clustering on patient utilization features. Which package from the chapter most directly offers both capabilities?

A) statsmodels
B) scikit-learn ✓ Correct
C) matplotlib
D) IPython

Rationale: scikit-learn includes submodules for clustering and dimensionality reduction.

Q81.
A new analyst downloads the book materials as a zip file and tries to run examples immediately, but file paths fail. According to the chapter, what should the analyst do first?

A) Convert all notebooks to PDFs
B) Fully extract the zip contents and navigate to that directory from the terminal ✓ Correct
C) Move only the datasets folder to the desktop
D) Install a different Python interpreter

Rationale: The chapter states that zip contents must be fully extracted and the user should navigate to that directory before running examples.

Q82.
A student is copying examples from the book and wants output formatting to match the printed examples more closely. Which action is recommended?

A) Use only default NumPy and pandas display options
B) Execute the provided pandas and NumPy display configuration code before running examples ✓ Correct
C) Disable IPython output prompts
D) Switch from Python 3 to Python 2

Rationale: The chapter provides pandas and NumPy option settings to match the book's displayed output formatting.

Q83.
An instructor sees a student writing from numpy import * in course materials. Based on the chapter's import conventions, what is the best correction?

A) Use import numpy as np ✓ Correct
B) Use import numpy as pandas
C) Use from numpy import ndarray only
D) Use import * only in production systems

Rationale: The chapter recommends the conventional alias import numpy as np and notes that importing everything from a large package is bad practice.

Python for Data Analysis 3rd edition McKinney Test Bank

Q84.
A student sees code using plt.plot(...) in a notebook and asks what plt refers to. Which import convention explains this?

A) import pandas as plt
B) import matplotlib.pyplot as plt ✓ Correct
C) import patsy as plt
D) import scipy.stats as plt

Rationale: plt is the standard alias for matplotlib.pyplot.

Q85.
A health data analyst with no prior Python experience wants to start the book efficiently. Which advice best follows the chapter's navigation guidance?

A) Skip directly to advanced modeling chapters
B) Spend time in Chapters 2 and 3 covering Python basics, IPython, and Jupyter ✓ Correct
C) Start with Appendix A only
D) Focus first on TensorFlow and PyTorch resources

Rationale: The chapter recommends that readers new to Python spend time in Chapters 2 and 3 because they are prerequisite knowledge.

Q86.
A manager asks a new analyst to classify the task of reading claims data from multiple file formats and databases into the book's broad workflow groups. Which group best fits this task?

A) Preparation
B) Transformation
C) Interacting with the outside world ✓ Correct
D) Presentation

Rationale: Reading and writing various file formats and data stores is categorized as interacting with the outside world.

Q87.
A data quality specialist standardizes dates, handles missing values, combines files, and reshapes a table before analysis. Which broad task group best describes this work?

A) Preparation ✓ Correct
B) Modeling and computation
C) Presentation
D) Community and conferences

Rationale: Cleaning, combining, normalizing, reshaping, and transforming data for analysis are part of preparation.

Python for Data Analysis 3rd edition McKinney Test Bank

Q88.

A population health analyst aggregates a large table by clinic and month to derive a new summary dataset. Which broad task group does this most closely represent?

A) Presentation

B) Transformation ✓ Correct

C) Installation and setup

D) Community support

Rationale: Applying mathematical or statistical operations to derive new datasets, such as aggregation by groups, is transformation.

Q89.

A researcher connects a curated dataset to a machine learning algorithm to estimate future utilization. Which broad task group best matches this activity?

A) Preparation

B) Presentation

C) Modeling and computation ✓ Correct

D) Interacting with the outside world

Rationale: Connecting data to statistical models, machine learning algorithms, or other computational tools is modeling and computation.

Q90.

A quality team creates static charts and narrative summaries for executive review after completing analysis. Which broad task group best describes this final step?

A) Presentation ✓ Correct

B) Transformation

C) Preparation

D) Optimization

Rationale: Creating graphical visualizations or textual summaries is the presentation stage.

Q91.

A student wants to run the entire set of book examples and cannot access GitHub from the workplace network. What alternative source is specifically mentioned?

A) Docker Hub

B) Gitee mirror ✓ Correct

C) PyPI archive

D) Kaggle repository

Rationale: The chapter notes that a mirror on Gitee may be used if GitHub is inaccessible.

Python for Data Analysis 3rd edition McKinney Test Bank

Q92.
A nurse analyst needs help with pandas-related questions from the community. Which mailing list is the most directly relevant according to the chapter?

A) numpy-discussion
B) pydata ✓ Correct
C) scipy-user
D) EuroPython announcements

Rationale: The pydata mailing list is identified for questions related to Python for data analysis and pandas.

Q93.
A statistician working with regression and econometric methods in Python wants a community list specifically relevant to statsmodels questions. Which list is the best fit?

A) pystatsmodels ✓ Correct
B) numpy-discussion
C) pycon
D) scikit-image

Rationale: pystatsmodels is listed for statsmodels or pandas-related questions.

Q94.
A clinician interested in broad data science and analytics networking wants a conference series targeted specifically at data analysis and data science use cases. Which conference is the best match?

A) EuroSciPy only
B) PyData ✓ Correct
C) PyCon only
D) Komodo IDE Summit

Rationale: PyData is described as a worldwide series of regional conferences targeted at data science and data analysis use cases.

Q95.
A student on Windows is copying the multiline conda install command from the chapter but receives syntax errors because of the line continuation character. What adjustment is recommended?

A) Replace the backslash with a semicolon
B) Replace the backslash with a carat ✓ Correct
C) Replace the backslash with a colon
D) Delete all line continuations and package names

Rationale: On Windows, the chapter says to substitute a carat (^) for the backslash line continuation used on Linux and macOS.

Python for Data Analysis 3rd edition McKinney Test Bank

Q96.

A project lead says that because the book is titled around data analysis, it will mainly teach advanced statistical methodology and machine learning theory. Which correction best reflects the chapter's stated focus?

A) The book focuses primarily on website development frameworks

B) The book focuses primarily on Python programming, libraries, and tools for data analysis rather than data analysis methodology ✓ Correct

C) The book focuses exclusively on deep learning architectures

D) The book focuses only on conference participation and community resources

Rationale: The chapter explicitly states that the focus is on Python programming, libraries, and tools rather than data analysis methodology.

Q97.

A team has a collection of unstructured patient comments and wants to prepare them for sentiment analysis. Which approach from the chapter best illustrates converting such data into a structured form?

A) Store the comments only as PDFs

B) Transform the comments into a word frequency table ✓ Correct

C) Convert the comments into a matplotlib figure first

D) Compile the comments into C++ binaries

Rationale: The chapter gives the example of processing news articles into a word frequency table for sentiment analysis.

Q98.

A healthcare analyst receives evenly spaced daily census values for one unit and irregular follow-up timestamps for another dataset. According to the chapter, how should these be categorized?

A) Both are examples of time series data that can be structured ✓ Correct

B) Only evenly spaced data are considered structured

C) Only unevenly spaced data are considered structured

D) Neither can be analyzed in Python

Rationale: The chapter includes both evenly and unevenly spaced time series as forms of structured data.

Q99.

A department is selecting tools for exploratory coding and says an edit-compile-run workflow is essential for efficient data exploration. Which response best aligns with the chapter?

A) Interactive exploration is discouraged in Python

B) IPython supports an execute-explore workflow that is well suited to data analysis ✓ Correct

C) Only compiled languages support iterative debugging

D) Jupyter should be used only for final publication, not exploration

Rationale: The chapter emphasizes IPython's execute-explore workflow as beneficial for exploratory data analysis work.

Python for Data Analysis 3rd edition McKinney Test Bank

Q100.

A health system wants one language for prototyping analytics and building production data applications while still leveraging strong general-purpose software engineering capabilities. Which rationale most strongly supports choosing Python?

A) Python is limited to spreadsheets and cannot support production systems

B) Python combines improved open source data libraries with strength in general-purpose software engineering ✓ Correct

C) Python eliminates any need for lower-level code in all situations

D) Python is recommended only when statistical inference is unnecessary

Rationale: The chapter highlights Python's strong data libraries together with its suitability for general-purpose software engineering and production systems.

Q101.

A startup wants one language for exploratory analysis, prototyping, and deployment of data products to reduce handoff friction between researchers and engineers. Based on the chapter, which rationale best supports choosing Python?

A) Python eliminates the need for any compiled extensions in performance-sensitive code

B) Python is suitable for both prototyping and production, reducing the organizational cost of maintaining separate language stacks ✓ Correct

C) Python guarantees lower runtime latency than Java and C++ in production systems

D) Python is designed primarily for statistical inference rather than software engineering

Rationale: The chapter emphasizes Python as a way to address the two-language problem by serving both research and production needs.

Q102.

A team is processing millions of records and finds that 95% of runtime occurs in a small numerical kernel, while the rest of the code orchestrates I/O and workflow. Which strategy is most aligned with the chapter's discussion of Python as glue?

A) Rewrite the entire codebase in FORTRAN to avoid interpretation overhead everywhere

B) Keep orchestration in Python and optimize only the computational bottleneck in a lower-level language or with JIT tools ✓ Correct

C) Replace NumPy arrays with Python lists to simplify interoperability

D) Move all shell interaction out of Python because glue code dominates runtime

Rationale: The text recommends using Python for glue code while optimizing true bottlenecks with lower-level code or performance tools.

Q103.

An analyst must integrate decades-old FORTRAN routines for optimization into a modern data workflow without rewriting the legacy algorithms. Which evaluation most accurately reflects Python's role?

A) Python is unsuitable because legacy scientific libraries cannot interoperate with it

B) Python is effective because it can integrate C, C++, and FORTRAN code while coordinating higher-level workflow ✓ Correct

C) Python should be avoided because only pure Python code can be called from NumPy

D) Python is ideal only when all algorithms are implemented directly in pandas

Rationale: Python's success in scientific computing is partly attributed to its ease of integrating legacy C/C++/FORTRAN code.

Python for Data Analysis 3rd edition McKinney Test Bank

Q104.

A quantitative trading firm needs single-process, highly CPU-bound multithreaded execution with extremely low latency. Which recommendation is most defensible from the chapter?

A) Use pure Python threads because the GIL improves CPU-bound parallelism

B) Prefer Python because programmer productivity always outweighs runtime constraints

C) Consider a lower-level language or native extensions because Python may be less suitable for this workload ✓ Correct

D) Use Jupyter notebooks as the main optimization mechanism for low-latency execution

Rationale: The chapter notes Python may be less suitable for very low-latency and CPU-bound multithreaded applications due to performance limits and the GIL.

Q105.

A data scientist argues that Python cannot execute truly parallel code under any circumstances because of the GIL. Which critique is most accurate?

A) Correct, because the GIL prevents all forms of parallel execution in every Python-related context

B) Incorrect, because native C/C++ extensions using multithreading can run in parallel when they do not frequently interact with Python objects ✓ Correct

C) Correct, because NumPy and SciPy are implemented entirely in Python bytecode

D) Incorrect, because the GIL affects only I/O-bound threads and never CPU-bound threads

Rationale: The chapter states that Python C extensions can execute parallel code without being impacted by the GIL if they avoid frequent Python-object interaction.

Q106.

A research group is selecting a core structure for large numerical datasets that must be passed efficiently among libraries without repeated copying. Which choice best fits the chapter's reasoning?

A) Python tuples because they preserve labels and relational metadata

B) NumPy ndarray because it is an efficient container for numerical data and a common interoperability target ✓ Correct

C) pandas Series because it is always more memory-efficient than ndarray for raw numerical arrays

D) Markdown tables because they are portable across notebook environments

Rationale: NumPy arrays are described as efficient numerical containers widely used for passing data between algorithms and libraries.

Q107.

A team receives sales data, sensor time series, and customer records linked across multiple tables. Which conclusion best synthesizes the chapter's treatment of structured data?

A) Only spreadsheet-like tables count as structured data

B) Structured data includes tabular data, multidimensional arrays, related tables, and time series ✓ Correct

C) Structured data excludes unevenly spaced observations

D) Structured data must originate from a relational database

Rationale: The chapter defines structured data broadly to include tabular data, arrays, interrelated tables, and evenly or unevenly spaced time series.

Q108.

A manager says unstructured text cannot be used in the book's data workflow because the focus is on structured data. Which response is best supported?

A) Agree, because text data cannot be transformed for analysis in Python

B) Disagree, because text can often be converted into structured features such as a word-frequency table ✓ Correct

C) Agree, because feature extraction belongs only to machine learning libraries, not data manipulation

D) Disagree, because pandas stores raw articles as regression coefficients

Rationale: The chapter notes that many datasets can be transformed into structured form or have features extracted into structured representations.

Q109.

An instructor wants students to focus on data wrangling rather than advanced modeling. Which library should receive the greatest emphasis according to the chapter's priorities?

A) pandas ✓ Correct

B) TensorFlow

C) PyTorch

D) statsmodels

Rationale: The chapter identifies pandas as a primary focus because manipulation, preparation, and cleaning are central data analysis skills.

Q110.

A developer is comparing pandas with spreadsheets and SQL-based systems. Which evaluation best captures pandas' design position?

A) It replaces NumPy by avoiding array computation entirely

B) It blends NumPy-style array computing with spreadsheet and relational-database style data manipulation ✓ Correct

C) It is primarily a plotting library with limited data structures

D) It is intended only for non-tabular machine learning inputs

Rationale: The chapter describes pandas as combining NumPy's array ideas with spreadsheet and SQL-like data manipulation capabilities.

Q111.

A data integration pipeline repeatedly merges datasets from different sources and has produced subtle errors from mismatched row labels. Which pandas design feature most directly addresses this risk?

A) Automatic or explicit data alignment on labeled axes ✓ Correct

B) Built-in support for GPU training

C) Mandatory positional indexing for all operations

D) Exclusive use of unlabeled matrices

Rationale: The chapter highlights labeled axes with automatic or explicit alignment as a safeguard against misaligned data errors.

Python for Data Analysis 3rd edition McKinney Test Bank

Q112.

A team needs one structure that can handle both time series and non-time-series business data without switching tools. Which historical motivation for pandas is most relevant?

A) It was designed only for image tensors and sparse graphs

B) It was created to unify handling of time-indexed and non-time-indexed structured data ✓ Correct

C) It was built mainly to replace Jupyter notebooks for reporting

D) It was intended to avoid all relational operations such as merges

Rationale: One of pandas' original requirements was that the same data structures handle both time series and non-time-series data.

Q113.

An analyst's primary goal is publication-quality two-dimensional graphics that integrate reasonably well with the Python data ecosystem. Which tool is the safest default choice from the chapter?

A) matplotlib ✓ Correct

B) patsy

C) Numba

D) conda-forge

Rationale: The chapter identifies matplotlib as the most popular and safe default library for 2D plotting and publication-quality figures.

Q114.

A student alternates between testing snippets, exploring outputs, and debugging code interactively while also needing shell and filesystem access in one place. Which environment best matches this workflow?

A) A compiled-only C++ toolchain

B) IPython, optionally through Jupyter ✓ Correct

C) A spreadsheet without scripting support

D) A SQL client with no Python integration

Rationale: IPython is presented as supporting an execute-explore workflow with integrated shell and filesystem access, and it can also be used via Jupyter.

Q115.

A faculty member wants a notebook environment that supports rich documents with code, Markdown, and HTML across many programming languages. Which conclusion is best supported?

A) Only IPython shell provides language-agnostic notebooks

B) Jupyter was created as a broader language-agnostic interactive computing project and supports rich notebook documents ✓ Correct

C) matplotlib is the preferred document-authoring system for multilingual notebooks

D) pandas DataFrame is the main notebook kernel for over 40 languages

Rationale: The chapter explains that Jupyter broadened the original IPython notebook into a language-agnostic system supporting many languages and rich text formats.

Q116.

A scientist needs specialized routines for numerical integration, sparse matrices, and statistical distributions in addition to NumPy. Which library is the most appropriate recommendation?

A) SciPy ✓ Correct
B) seaborn
C) beautifulsoup4
D) openpyxl

Rationale: SciPy is described as a collection of packages for foundational scientific computing tasks including integration, sparse matrices, and statistics.

Q117.

A machine learning engineer wants a Python toolkit centered on predictive models, preprocessing, model selection, and clustering. Which package best fits?

A) statsmodels
B) scikit-learn ✓ Correct
C) matplotlib
D) html5lib

Rationale: The chapter describes scikit-learn as the premier general-purpose machine learning toolkit with modules for prediction, preprocessing, clustering, and model selection.

Q118.

A policy analyst must estimate coefficients, report p-values, and discuss uncertainty rather than optimize predictive accuracy alone. Which package is the better choice?

A) scikit-learn because it is more inference-focused
B) statsmodels because it emphasizes classical statistics, econometrics, and parameter uncertainty ✓ Correct
C) NumPy because it automatically reports inferential statistics
D) Jupyter because notebook cells compute p-values by default

Rationale: The chapter contrasts statsmodels as inference-focused with scikit-learn as more prediction-focused.

Q119.

A student is trying to decide whether to begin with TensorFlow/PyTorch or first build stronger data-wrangling skills. Which recommendation best reflects the chapter's sequencing logic?

A) Start with deep learning frameworks because foundational wrangling can be skipped
B) Build a foundation in general-purpose Python data wrangling first, then move to more advanced specialized resources ✓ Correct
C) Avoid pandas until after mastering neural network libraries
D) Use only R before learning Python data tools

Rationale: The chapter recommends using the book to build core Python data wrangling skills before advancing to more specialized tools like TensorFlow or PyTorch.

Python for Data Analysis 3rd edition McKinney Test Bank

Q120.

An enterprise team wants a reproducible cross-platform setup and plans to use Miniconda with community-maintained packages. Which setup decision is explicitly recommended?

A) Use pip as the default package source and conda only as backup

B) Configure conda-forge as the default channel with strict channel priority ✓ Correct

C) Install packages globally without environments to reduce duplication

D) Prefer OS package managers over Miniconda for newest package versions

Rationale: The installation instructions explicitly recommend adding conda-forge and setting channel priority to strict.

Q121.

A learner creates a conda environment once and assumes it will remain active in all future terminal sessions. Which evaluation is correct?

A) Correct, because conda environments persist as active across terminal restarts

B) Incorrect, because the environment must be activated each time a new terminal is opened ✓ Correct

C) Correct, but only on Windows

D) Incorrect, because conda environments cannot be reactivated after creation

Rationale: The chapter states that you must use conda activate each time you open a new terminal.

Q122.

A user installed Miniconda packages with conda and now plans to upgrade those same packages with pip for convenience. Which recommendation best reflects the chapter's caution?

A) Proceed, because mixing package managers improves dependency resolution

B) Avoid doing so, because updating conda-installed packages with pip can create environment problems ✓ Correct

C) Use pip first and then conda update on the same package to ensure consistency

D) Mixing tools is required for Jupyter to run correctly

Rationale: The chapter cautions against updating packages installed by one manager using the other because it can break environments.

Q123.

A Linux administrator argues that system package managers are always preferable to Miniconda for this book because they are native to the OS. Which counterargument is most aligned with the text?

A) Miniconda is recommended because it is more reproducible across distributions and simplifies upgrading to newer package versions ✓ Correct

B) System package managers are rejected because they cannot install Python at all

C) Miniconda is required only for Windows, not Linux

D) Linux users must compile all packages manually from source

Rationale: The chapter recommends Miniconda on Linux for reproducibility across distributions and easier access to current package versions.

Python for Data Analysis 3rd edition McKinney Test Bank

Q124.

A macOS user with an Apple Silicon machine downloads the x86_64 installer because it is more familiar. Based on the chapter, what is the best evaluation?

A) Appropriate, because Apple Silicon Macs should always use x86_64 installers

B) Potentially inappropriate, because Apple Silicon systems should generally use the arm64 Miniconda installer ✓ Correct

C) Required, because arm64 installers are intended only for Linux

D) Irrelevant, because Miniconda installer architecture never matters

Rationale: The macOS instructions distinguish arm64 installers for Apple Silicon from x86_64 installers for Intel Macs.

Q125.

A Windows user declines adding Miniconda to PATH and then cannot launch the intended Python from a regular command prompt. Which action best follows the chapter's guidance?

A) Use the Start menu shortcut such as the Anaconda Prompt installed with Miniconda ✓ Correct

B) Reinstall Python from source immediately

C) Use only JupyterHub in a browser because local shells are unsupported

D) Delete other Python installations to force PATH resolution

Rationale: If Miniconda is not added to PATH, the chapter advises using the installed Start menu shortcut to access it.

Q126.

A student needs only the core tools used throughout the book right after creating the pydata-book environment. Which package set most directly matches the initial installation step?

A) pandas, jupyter, matplotlib ✓ Correct

B) scikit-learn, statsmodels, seaborn

C) numpy, scipy, tensorflow

D) sqlalchemy, requests, pyarrow

Rationale: The chapter's initial conda install command for the book installs pandas, jupyter, and matplotlib.

Q127.

A learner sees a package unavailable through conda install in the Miniconda environment. Which decision rule best matches the chapter?

A) Always compile from source before trying pip

B) Prefer conda install, but use pip install if the package is unavailable through conda ✓ Correct

C) Use pip for all upgrades and conda only for environment creation

D) Switch to the system Python because conda cannot manage missing packages

Rationale: The chapter explicitly recommends preferring conda and falling back to pip only when needed.

Python for Data Analysis 3rd edition McKinney Test Bank

Q128.

A data analyst asks whether the book's recommended workflow is centered on a full IDE or on interactive tools plus an editor. Which answer is most accurate?

A) The author strongly recommends only heavyweight IDEs for all tasks

B) The recommended baseline workflow is IPython plus a text editor, with IDEs as optional alternatives ✓ Correct

C) The book discourages interactive exploration during development

D) The preferred workflow is spreadsheet-first, Python-second

Rationale: The chapter states that the author's standard environment is typically IPython plus a text editor, while acknowledging IDE alternatives.

Q129.

A novice with no prior Python experience wants to jump directly into later pandas chapters. Which advice best reflects the chapter's navigation strategy?

A) Skip early chapters because Python syntax is irrelevant to data analysis

B) Spend time in Chapters 2 and 3 first, since Python language features and IPython/Jupyter are prerequisites ✓ Correct

C) Begin with Appendix A because advanced NumPy replaces the need for Python basics

D) Start with conference resources before reading technical chapters

Rationale: The chapter advises beginners to study Chapters 2 and 3 because they provide prerequisite Python and IPython/Jupyter knowledge.

Q130.

A curriculum designer wants to classify tasks such as cleaning, combining, normalizing, reshaping, and slicing data. Under which broad group do these belong?

A) Presentation

B) Preparation ✓ Correct

C) Modeling and computation

D) Interacting with the outside world

Rationale: The chapter lists cleaning, munging, combining, normalizing, reshaping, slicing, and transforming under Preparation.

Q131.

A team aggregates a large table by group variables to derive a summary dataset for subsequent analysis. According to the chapter's task taxonomy, this is primarily an example of:

A) Transformation ✓ Correct

B) Presentation

C) Interacting with the outside world

D) Installation and setup

Rationale: The chapter defines Transformation as applying mathematical and statistical operations to derive new datasets, including aggregation by groups.

Python for Data Analysis 3rd edition McKinney Test Bank

Q132.

A researcher is linking cleaned data to machine learning algorithms and statistical models. Which broad task group best fits this work?

A) Preparation
B) Presentation
C) Modeling and computation ✓ Correct
D) Community and conferences

Rationale: The chapter identifies connecting data to statistical models and machine learning algorithms as Modeling and computation.

Q133.

A user wants to generate static charts and textual summaries for stakeholders after completing analysis. Which broad task group is most appropriate?

A) Presentation ✓ Correct
B) Transformation
C) Preparation
D) Import conventions

Rationale: The chapter classifies graphical visualizations and textual summaries under Presentation.

Q134.

A student expects exact visual agreement with every printed example in the book but notices different numeric precision and display width. Which action is most justified?

A) Assume the examples are wrong because NumPy and pandas output can never be configured
B) Apply the display and print option settings provided in the chapter before running examples ✓ Correct
C) Reinstall Python 3.9 instead of 3.10
D) Switch all imports to wildcard syntax

Rationale: The chapter provides specific pandas and NumPy display settings to match the book's printed output more closely.

Q135.

A learner downloads the repository as a zip file and tries to run examples without extracting it or changing into the project directory. Which evaluation is most accurate?

A) This is fine because Jupyter automatically finds all datasets inside compressed archives
B) This may cause problems because the zip contents should be fully extracted and the terminal should be navigated to that directory ✓ Correct
C) Extraction is needed only on Windows, not Linux or macOS
D) The datasets are embedded in pandas, so local files are unnecessary

Rationale: The chapter instructs users who download a zip file to fully extract it and navigate to that directory before running examples.

Python for Data Analysis 3rd edition McKinney Test Bank

Q136.

A reader finds a likely omission in the example materials and wants the most appropriate route for reporting reproducibility issues. Which option best reflects the chapter?

A) Post only on a random social media site because official channels are not mentioned

B) Email the author about repository omissions and use the O'Reilly errata page to report book errors ✓ Correct

C) Report all issues exclusively to NumPy maintainers

D) Open a ticket with the operating system vendor

Rationale: The chapter provides the author's email for repository issues and recommends the O'Reilly errata page for book errors.

Q137.

A code reviewer sees from numpy import * in a shared analytics codebase and asks whether this aligns with the chapter's import conventions. Which judgment is best supported?

A) It is recommended because wildcard imports improve clarity in large packages

B) It is discouraged as bad practice; standard aliases like np and pd are preferred ✓ Correct

C) It is required for compatibility with Jupyter notebooks

D) It is preferred only for matplotlib, not NumPy

Rationale: The chapter states that importing everything from a large package like NumPy is considered bad practice and shows standard aliases instead.

Q138.

A new analyst sees np.arange in example code and is unsure what np refers to. Which interpretation is correct under the community's import conventions?

A) np is a variable containing notebook output history

B) np is the customary alias for the NumPy module ✓ Correct

C) np is a pandas accessor for numeric precision

D) np is the SciPy namespace for nonlinear programming

Rationale: The chapter lists import numpy as np as a standard convention, so np.arange refers to NumPy's arange function.

Q139.

A colleague claims pandas is mainly a machine learning package because it has labeled arrays and DataFrames. Which rebuttal best reflects the chapter?

A) pandas is primarily focused on structured data manipulation, preparation, cleaning, indexing, reshaping, and relational operations ✓ Correct

B) pandas exists mainly for GPU-accelerated neural network training

C) pandas is a plotting backend for matplotlib with no data structures

D) pandas is identical in purpose to scikit-learn model selection tools

Rationale: The chapter emphasizes pandas as a high-level library for structured/tabular data manipulation rather than a machine learning toolkit.

Q140.

A team must choose between scikit-learn and statsmodels for a project where the main deliverable is an interpretable econometric report with uncertainty estimates. Which recommendation is strongest?

A) Use scikit-learn because it is more focused on p-values and frequentist inference

B) Use statsmodels because it is oriented toward classical statistics, econometrics, and inference outputs ✓ Correct

C) Use matplotlib because publication-quality plots replace model interpretation

D) Use Jupyter only, since kernels provide inferential estimates automatically

Rationale: The chapter specifically positions statsmodels as more focused on statistical inference and econometrics than scikit-learn.

Q141.

An instructor asks students to compare IPython's workflow with the traditional edit-compile-run cycle common in other languages. Which interpretation best captures the chapter's contrast?

A) IPython promotes an execute-explore workflow suited to iterative analysis and debugging ✓ Correct

B) IPython requires a compile step before any code can be tested

C) IPython is intended only for final report generation, not exploration

D) IPython replaces the need for files, editors, and debugging entirely

Rationale: The chapter contrasts IPython's execute-explore style with traditional edit-compile-run workflows.

Q142.

A systems architect is deciding whether Python should be the sole implementation language for every part of a data platform, including all high-performance kernels. Which evaluation is most nuanced according to the chapter?

A) Yes, because Python always matches compiled languages in speed

B) No, because Python is useless for production systems

C) Often yes for much of the system, but some computational kernels may still benefit from lower-level languages or JIT acceleration ✓ Correct

D) No, because NumPy cannot interoperate with native code

Rationale: The chapter presents Python as suitable for many production uses while acknowledging that some bottlenecks may require C/C++ or JIT tools like Numba.

Q143.

A data engineering lead argues that because Python is slower than compiled languages, it should never be used in analytics systems. Which response best evaluates that claim?

A) Agree, because runtime speed is always more valuable than developer productivity

B) Disagree, because many users accept slower execution when programmer time is more valuable than CPU time ✓ Correct

C) Agree, because interpreted languages cannot support serious software

D) Disagree, because Python is actually compiled to machine code by default

Rationale: The chapter notes that many users are willing to trade some runtime performance for higher programmer productivity.

Python for Data Analysis 3rd edition McKinney Test Bank

Q144.

A learner wants the package most closely associated with formula-based model specification inspired by R's formula system in support of statistical modeling workflows. Which project is most directly connected to that capability?

A) Patsy ✓ Correct

B) matplotlib

C) Numba

D) openpyxl

Rationale: The chapter states that Nathaniel Smith developed Patsy to provide a formula or model specification framework for statsmodels inspired by R.

Q145.

A student is comparing NumPy and SciPy and asks whether either alone forms the full computational basis emphasized for traditional scientific computing. Which answer best synthesizes the chapter?

A) Only SciPy matters because NumPy is mainly for visualization

B) Only NumPy matters because SciPy duplicates no additional functionality

C) Together NumPy and SciPy provide a mature computational foundation for many traditional scientific applications ✓ Correct

D) Neither is relevant because pandas replaces both for scientific computing

Rationale: The chapter explicitly states that NumPy and SciPy together form a reasonably complete and mature computational foundation.

Q146.

A conference organizer wants to direct participants interested specifically in data science and data analysis use cases rather than general Python or broad scientific computing. Which conference series is the best fit?

A) EuroPython

B) PyData ✓ Correct

C) PyCon only

D) EuroSciPy

Rationale: The chapter identifies PyData as a worldwide series of regional conferences targeted at data science and data analysis use cases.

Q147.

A practitioner has a question about pandas data analysis workflows and wants a community venue most directly aligned with that topic. Which mailing list is the best initial choice?

A) numpy-discussion

B) pydata ✓ Correct

C) scipy-user

D) scikit-learn@python.org

Rationale: The chapter lists pydata as a Google Group for questions related to Python for data analysis and pandas.

Python for Data Analysis 3rd edition McKinney Test Bank

Q148.

A faculty member wants students to infer the book's pedagogical scope. Which statement best captures that scope at an evaluative level?

A) The book primarily teaches advanced machine learning methodology in Python

B) The book focuses on Python programming, libraries, and tools needed for data analysis rather than on data analysis methodology itself ✓ Correct

C) The book is mainly about relational database administration

D) The book is centered on web development frameworks such as Django and Rails

Rationale: The chapter explicitly distinguishes its focus on Python tools and programming from broader data analysis methodology.

Q149.

An analyst must decide whether a collection of news articles belongs within the book's practical notion of analyzable data. Which conclusion best integrates the chapter's reasoning?

A) It does not, because only native numeric arrays can be analyzed

B) It can, if transformed into structured features such as a word-frequency table for downstream analysis ✓ Correct

C) It cannot, because sentiment analysis requires only spreadsheets

D) It belongs only if stored first as a SQL primary key

Rationale: The chapter uses news articles transformed into a word-frequency table as an example of extracting structured features from unstructured data.

Q150.

A team is choosing a default interactive environment for a reproducible training program. They need code execution, exploratory iteration, notebook support, and easy transition between shell-based and document-based workflows. Which recommendation is most defensible from the chapter as a whole?

A) Adopt IPython and Jupyter as complementary tools for interactive computing and exploratory development ✓ Correct

B) Use only a commercial IDE because notebooks hinder reproducibility

C) Avoid shell-based tools and rely exclusively on static scripts

D) Standardize on matplotlib because plotting libraries also manage environments

Rationale: The chapter presents IPython and Jupyter as central, complementary tools for interactive, exploratory, and notebook-based Python work.

Python for Data Analysis 3rd edition McKinney Test Bank