PCAD-31-0X

Block 1: Data Acquisition and Pre-Processing (33% of total exam)
Block 2: Programming and Database Skills (29% of total exam)
Block 3: Statistical Analysis (9% of total exam)
Block 4: Data Analysis and Modeling (16% of total exam)
Block 5: Data Communication and Visualization (13% of total exam)

Block 1: Data Acquisition and Pre-Processing (33% of total exam)

Objectives covered by the block (15 exam items)

Data Collection, Integration, and Storage

- **Objective 1.1.1** – Understand different data collection methods and their roles in decision-making and research.
    - Explore different techniques: Surveys, interviews, web scraping.
    - Discuss representative sampling, challenges in data collection, and differences between qualitative and quantitative research.
    - Examine legal and ethical considerations in data collection.
    - Explain the importance of data anonymization in maintaining privacy and confidentiality, particularly with personally identifiable information (PII).
    - Investigate the impact of data collection on business strategy formation, market research accuracy, risk assessment, policy-making, and business decisions.
- **Objective 1.1.2** – Explain the data gathering process and various data sources.
    - Explain the process and methodologies of data collection, including survey design, audience selection, and structured interviews.
- **Objective 1.1.3** – Aggregate data from multiple sources and integrate them into datasets.
    - Explain techniques for combining data from various sources, such as databases, APIs, and file-based storage.
    - Address challenges in data aggregation, including data format disparities and alignment issues.
    - Understand the importance of data consistency and accuracy in aggregated datasets.
- **Objective 1.1.4** – Explain various data storage solutions.
    - Understand various data storage methods and their appropriate applications.

- Distinguish between the concepts of data warehouses, data lakes, and file-based storage options like CSV and Excel.
- Explain the concepts of cloud storage solutions and their growing role in data management.

Data Cleaning and Standardization

- **Objective 1.2.1** – Understand structured and unstructured data and their implications in data analysis.

  - Recognize the characteristics of structured data, such as databases and spreadsheets, and their straightforward use in analysis.
  - Understand unstructured data, including text, images, and videos, and the additional processing required for analysis.
  - Explore how the data structure impacts data storage, retrieval, and analytical methods.

- **Objective 1.2.2** – Identify, rectify, or remove erroneous data.

  - Identify data errors and inconsistencies through various diagnostic methods.
  - Address missing, inaccurate, or misleading information.
  - Tackle specific data quality issues: numerical data problems, duplicate records, invalid data entries, and missing values.
  - Explain different types of missingness (MCAR, MAR, MNAR), and their implications for data analysis.
  - Explore various techniques for dealing with missing data, including data imputation methods.
  - Understand the implications of data correction or removal on overall data integrity and analysis outcomes.
  - Explain the importance of data collection in the context of outlier detection.
  - Explain why high-quality data is crucial for accurate outlier detection.
  - Explain how different data types (numerical, categorical) may influence outlier detection strategies.

- **Objective 1.2.3** – Understand data normalization and scaling.

  - Understand the necessity of data normalization to bring different variables onto a similar scale for comparative analysis.
  - Understand various scaling methods like Min-Max scaling and Z-score normalization.
  - Explain encoding categorical variables for quantitative analysis, including one-hot encoding and label encoding methods.
  - Explain the pros and cons of data reduction (reduce the number of variables under consideration or simplify the models vs loss of data explainability).
  - Explain methods for handling outliers, including detection and treatment techniques to ensure data quality.

    - Understand the importance of data format standardization across different datasets for consistency, especially when dealing with date-time formats and numerical values.

- **Objective 1.2.4** – Apply data cleaning and standardization techniques.

    - Perform data imputation techniques, string manipulation, data format standardization, boolean normalization, string case normalization, and string-to-number conversions.
    - Discuss the pros and cons of imputation vs. exclusion and their impact on the reliability and validity of the analysis.
    - Explain the concept of One-Hot Encoding and its application in transforming categorical variables into a binary format, and preparing data for machine learning algorithms.
    - Explain the concept of bucketization and its application in transforming continuous variables into categorical variables.

Data Validation and Integrity

- **Objective 1.3.1** – Execute and understand basic data validation methods.

    - Perform type, range, and cross-reference checks.

- **Objective 1.3.2** – Establish and maintain data integrity through clear validation rules.

    - Understand the concept of data integrity and its importance in maintaining reliable and accurate databases.
    - Apply clear validation rules that enforce the correctness and consistency of data.

Data Preparation Techniques

- **Objective 1.4.1** – Understand File Formats in Data Acquisition.

    - Explain the roles and characteristics of common data file formats: CSV for tabular data, JSON for structured data, XML for hieXrarchically organized data, and TXT for unstructured text.
    - Understand basic methods for importing and exporting these file types in data analysis tools, focusing on practical applications.

- **Objective 1.4.2** – Access, manage, and effectively utilize datasets.

    - Understand the basics of accessing datasets from various sources like local files, databases, and online repositories.
    - Understand the principles of data management, including organizing, sorting, and filtering data in preparation for analysis.

- **Objective 1.4.3** – Extract data from various sources.

    - Explain fundamental techniques for extracting data from various sources, emphasizing methods to retrieve and collate data from databases, APIs, and online services.
    - Understand basic challenges and considerations in data extraction, such as data compatibility and integrity.

- **Objective 1.4.4** – Enhance data readability and format in spreadsheets.

    - Improve the readability and usability of data in spreadsheets, focusing on layout adjustments, formatting best practices, and basic formula applications.

- **Objective 1.4.5** – Prepare, adapt, and pre-process data for analysis.

    - Understand the importance of the surrounding context, objectives and stakeholder expectations to guide the preparation steps.
    - Understand basic concepts of data pre-processing, including sorting, filtering, and preparing data sets for analytical work.
    - Discuss the importance of proper data formatting for analysis, such as ensuring consistency in date-time formats and aligning data structures.
    - Introduce concepts of dataset structuring, including the basics of transforming data into a format suitable for analysis (e.g., wide vs. long formats).
    - Explain the concept of splitting data into training and testing sets, particularly for machine learning projects, emphasizing the importance of this step for model validation.
    - Understand the impact of outlier management on data quality in preprocessing.

Block 2: Programming and Database Skills (29% of total exam)

Objectives covered by the block (13 exam items)

Python Proficiency

- **Objective 2.1.1** – Apply Python syntax and control structures to solve data-related problems.
    - Accurately use basic Python syntax for variables, scopes, and data types.
    - Implement control structures like loops and conditionals to manage data flow.
- **Objective 2.1.2** – Analyze and create Python functions.

    - Design functions with clear purpose, using both indexed and keyword arguments.
    - Differentiate between optional and required arguments and apply them effectively.

- **Objective 2.1.3** – Evaluate and navigate the Python Data Science ecosystem.
    - Identify key Python libraries and tools essential for data science tasks.
    - Critically assess the suitability of various Python resources for different data analysis scenarios.

- **Objective 2.1.4** – Organize and manipulate data using Python's core data structures.
    - Effectively use tuples, sets, lists, dictionaries, and strings for data organization and manipulation.
    - Solve complex data handling tasks by choosing appropriate data structures.

- **Objective 2.1.5** – Explain and implement Python scripting best practices.
    - Understand and apply PEP 8 guidelines for Python coding style.
    - Comprehend and utilize PEP 257 for effective docstring conventions to enhance code documentation.

Module Management and Exception Handling

- **Objective 2.2.1** – Import modules and manage Python packages using PIP.
    - Apply different types of module imports (standard imports, selective imports, aliasing).
    - Understand importing modules from different sources (Python Standard Library, via package managers like PIP, and from locally developed modules/packages).
    - Identify and import necessary Python modules for specific tasks, understanding the functionality and purpose of each.
    - Demonstrate proficiency in managing Python packages using PIP, including installing, updating, and removing packages.

- **Objective 2.2.2** – Apply basic exception handling and maintain script robustness.
    - Implement basic exception handling techniques to manage and respond to errors in Python scripts.
    - Predict common errors in Python code and develop strategies to handle them effectively.
    - Interpret error messages to diagnose and resolve issues, enhancing the robustness and reliability of Python scripts.

SQL for Data Analysts

- **Objective 2.3.1** – Perform SQL queries to retrieve and manipulate data.
    - Compose and execute SQL queries to extract data from database tables.

    - Apply SQL functions and clauses to manipulate and filter data effectively.
    - Construct and execute SQL queries using SELECT, FROM, JOINS (INNER, LEFT, RIGHT, FULL), WHERE, GROUP BY, HAVING, ORDER BY, and LIMIT.
    - Analyze data retrieval needs and apply appropriate clauses from the SFJWGHOL set to meet those requirements effectively.

- **Objective 2.3.2** – Execute fundamental SQL commands to create, read, update, and delete data in database tables.

    - Demonstrate the ability to use CRUD operations (Create, Read, Update, Delete) in SQL.
    - Construct SQL statements for data insertion, retrieval, updating, and deletion.

- **Objective 2.3.3** – Establish connections to databases using Python.

    - Understand and implement methods to establish database connections using Python libraries (e.g., sqlite3, pymysql).
    - Analyze and resolve common issues encountered while connecting Python scripts to databases.

- **Objective 2.3.4** – Execute parameterized SQL queries through Python to safely interact with databases.

    - Develop and execute parameterized SQL queries in Python to interact with databases securely.
    - Evaluate the advantages of parameterized queries in preventing SQL injection and maintaining data integrity.

- **Objective 2.3.5** – Understand, manage and convert SQL data types appropriately within Python scripts.

    - Identify and understand various SQL data types and their counterparts in Python.
    - Practice converting data types appropriately when transferring data between SQL databases and Python scripts.

- **Objective 2.3.6** – Understand essential database security concepts, including strategies to prevent SQL query injection.

    - Comprehend fundamental database security principles, including measures to prevent SQL injection attacks.
    - Assess and apply strategies for writing secure SQL queries within Python environments.

Block 3: Statistical Analysis (9% of total exam)

Objectives covered by the block (4 exam items)

Descriptive Statistics

- **Objective 3.1.1** – Understand and apply statistical measures in data analysis.
    - Understand and describe measures of central tendency and spread.
    - Identify fundamental statistical distributions (Gaussian, Uniform) and interpret their trends in various contexts (over time, univariate, bivariate, multivariate).
    - Apply confidence measures in statistical calculations to assess data reliability.
- **Objective 3.1.2** – Analyze and evaluate data relationships.
    - Analyze datasets to identify outliers and evaluate negative and positive correlations using Pearson's R coefficient.
    - Interpret and critically assess information presented in various types of plots and graphs, including Boxplots, Histograms, Scatterplots, Lineplots, and Correlation heatmaps.

Inferential Statistics

- **Objective 3.2.1** – Understand and apply bootstrapping for sampling distributions.
    - Understand the theoretical basis and statistical principles underlying bootstrapping.
    - Differentiate between discrete and continuous data types in the context of bootstrapping.
    - Recognize situations and data types where bootstrapping is an effective method for estimating sampling distributions.
    - Demonstrate proficiency in applying bootstrapping methods using Python to generate and analyze sampling distributions.
    - Analyze the reliability and validity of results obtained from bootstrapping in various statistical scenarios.
- **Objective 3.2.2** – Explain when and how to use linear and logistic regression.
    - Comprehend the theory, assumptions, and mathematical foundation of linear regression.
    - Explain the concepts, use cases, and statistical underpinnings of logistic regression.
    - Develop the ability to choose between linear and logistic regression based on the nature of the data and the research question.
    - Apply the concepts of discrete and continuous data in choosing and implementing linear and logistic regression models.

- Demonstrate the application of linear and logistic regression models on datasets using Python, including parameter estimation and model fitting.
- Accurately interpret the outcomes of regression analyses, including coefficients and model fit statistics.
- Identify limitations, assumptions, and potential biases in linear and logistic regression models and their impact on results.

Block 4: Data Analysis and Modeling (16% of total exam)

Objectives covered by the block (7 exam items)

Data Analysis with Pandas and NumPy

- **Objective 4.1.1** – Manage data effectively with Pandas.
  - Organize and clean data using Pandas' data manipulation tools (like filtering, sorting, and handling missing values).
  - Apply advanced data manipulation techniques such as merging, joining, and reshaping data frames.
- **Objective 4.1.2** – Understand and Utilize the Relationship Between DataFrame and Series in Pandas.
  - Explain the conceptual differences and connections between DataFrames and Series in Pandas.
  - Implement indexing methods and use vectorized functions for efficient data manipulation.
  - Practice locating data using .iloc and .loc methods, and analyze the outcomes to ensure accurate data retrieval and manipulation.
- **Objective 4.1.3** – Perform Array Operations and Differentiate Data Structures with NumPy.
  - Execute array operations using NumPy, including basic arithmetic, broadcasting, and aggregation functions.
  - Distinguish between arrays, lists, NDArrays, Series, and DataFrames, understanding their respective use cases and performance characteristics.
  - Analyze and compare the efficiency and suitability of these data structures for different types of data analysis tasks.
- **Objective 4.1.4** – Apply and Analyze Data Organization Techniques in Pandas and NumPy.
  - Apply methods for reshaping data, including subsetting and sorting, in Pandas.
  - Analyze datasets by grouping data using groupby and creating pivot/cross tables for enhanced data comprehension.
  - Compute and interpret descriptive statistics using Pandas and NumPy to extract meaningful insights from data.

Statistical Methods and Machine Learning

- **Objective 4.2.1** – Apply Python's descriptive statistics for dataset analysis.
    - Calculate and interpret key statistical measures such as mean, median, mode, variance, and standard deviation using Python.
    - Utilize Python libraries (like Pandas and NumPy) to generate and analyze descriptive statistics for real-world datasets.
- **Objective 4.2.2** – Recognize the importance of test datasets in model evaluation.
    - Understand the role of test datasets in validating the performance of machine learning models.
    - Demonstrate knowledge of proper test dataset selection and usage to ensure unbiased and accurate model evaluation.
- **Objective 4.2.3** – Analyze and Evaluate Supervised Learning Algorithms and Model Accuracy.
    - Analyze various supervised learning algorithms to understand their specific characteristics and applications.
    - Evaluate the concepts of overfitting and underfitting within these models, including a detailed explanation of the bias-variance tradeoff.
    - Assess the intrinsic tendencies of linear and logistic regression in relation to this tradeoff, and apply this understanding to prevent model accuracy issues.

Block 5: Data Communication and Visualization (13% of total exam)

Objectives covered by the block (6 exam items)

Data Visualization Techniques

- **Objective 5.1.1** – Demonstrate essential proficiency in data visualization with Matplotlib and Seaborn.
    - Utilize Matplotlib and Seaborn to create various types of plots, including Boxplots, Histograms, Scatterplots, Lineplots, and Correlation heatmaps.
    - Interpret the data and findings represented in these visualizations to gain deeper insights and communicate results effectively.
- **Objective 5.1.2** – Assess the pros and cons of different data representations.
    - Evaluate the suitability of various chart types for different types of data and analysis objectives.
    - Critically analyze the effectiveness of chosen visualizations in conveying the intended message or insight.
- **Objective 5.1.3** – Label, annotate, and test insights from data visualizations.

- Incorporate labels, titles, and annotations in visualizations to clarify and emphasize key insights.
- Utilize visual exploration to generate hypotheses and test insights from datasets.
- Practice making data-driven decisions based on the interpretation of visualized data.

- **Objective 5.1.4** – Improve the clarity and accuracy of data interpretation by managing display features such as colors, labels and legends.

    - Customize colors in plots to improve readability of a scatterplot.
    - Label axes and add titles to improve data readability.
    - Manipulate legend properties such as position, font size, and background color, to improve the esthetics and readability of data.

Effective Communication of Data Insights

- **Objective 5.2.1** – Tailor communication to different audience needs, and combine visualizations and text for clear data presentation.

    - Analyze the audience to understand their background, interests, and knowledge level.
    - Adapt communication style and content to meet the specific needs and expectations of diverse audiences.
    - Create presentations and reports that effectively convey data insights to both technical and non-technical stakeholders.
    - Integrate visualizations seamlessly into presentations and reports, aligning them with the narrative.
    - Use concise and informative text to complement visualizations, providing context and key takeaways.
    - Ensure visual and textual elements work harmoniously to enhance data clarity and understanding.
    - Avoid slide clutter and optimize slide content to maintain focus on key messages.
    - Craft a compelling data narrative that tells a story with data, highlighting insights and actionable takeaways.
    - Select an appropriate and consistent color palette for visualizations, ensuring clarity and accessibility.

- **Objective 5.2.2** – Summarize key findings and support claims with evidence and reasoning.

    - Understand the process of identifying and extracting key findings from data analysis.
    - Apply techniques to condense complex information into concise and meaningful summaries.
    - Prioritize and emphasize the most relevant insights based on context.
    - Explain the importance of backing assertions and conclusions with data-driven evidence and reasoning.

- Articulate the basis for claims and recommendations, demonstrating transparency in decision-making.
- Demonstrate proficiency in clearly presenting evidence to support claims and recommendations.