Python Data Analytics Coding challenge Questions

For Beginners & Early Intermediates

- **Use any of the 5 datasets given – or more if you wish.**
- **Goal:** Build a professional Python data analysis project by Day 56 using everything you've learnt so far.

CHALLENGE OVERVIEW

In this day-wise challenge, you'll move from basic Python operations to cleaning, analysing, and visualising data and generating insights. Each day's task builds on the previous one—so stay consistent.

Dataset Options:

- Titanic Dataset – Titanic Dataset
- Iris Dataset – Iris Dataset
- Penguins Dataset – Penguin Dataset
- HR-Employee-Attrition – HR-Employee-Attrition Dataset

NOTE: Choose one dataset from the provided dataset for the Python for data analysis module challenge. So that by the end of the challenge, you will have completed a portfolio-ready project.

Day 1 – NumPy Arrays & Operations

Task:

- Create 1D, 2D, 3D arrays.
- Perform addition, subtraction, multiplication, division.
- Use functions: `mean()`, `sum()`, `std()`, `reshape()`, `arange()`, `linspace()`.
- Find indices of max/min values.
  Project Connection: Begin numerical data exploration for the final dataset.

Day 2– Pandas Basics

Task:

- Import dataset into Pandas DataFrame.
- Explore series & columns.
- Add, update, delete columns/rows.
- Set custom index and sort.
  Project Connection: Ingest your chosen dataset and start basic cleaning & structuring.

Day 3 – Data Exploration & Wrangling

Task:

- Explore dataset: `head()`, `tail()`, `info()`, `describe()`, `value_counts()`.
- Detect & handle missing values: `dropna()`, `fillna()`.
- Merge, concat, join datasets.
- Group by and aggregate data.
  Project Connection: Perform data cleaning, initial wrangling, and summary statistics.

Day 5 – Matplotlib Visualization

Task:

- Line chart, scatter plot, bar chart, pie chart.
- Histogram & box plot to show distributions and outliers.
  Project Connection: Visualize trends, distributions, and outliers in your dataset.

Day 6 – Seaborn & Plotly Advanced Visualization

Task:

- Heatmap, pair plots, violin, swarm, density plots.
- Interactive Plotly charts (scatter, bar, line).
  Project Connection: Identify correlations and relationships; create interactive visuals.

Day 7– Data Cleaning & Stats

Task:

- Detect & remove duplicates.
- Handle inconsistent or missing data.
- Compute basic statistics: mean, median, mode, std, variance.
- Group by categories and summarize insights.
  Project Connection: Prepare a clean dataset and generate key statistical insights for the final project.

Day 8 – Introduction to Web Scraping & AI

Task:

- Use `requests` to fetch HTML content.
- Generate Python code via Gemini AI in Colab.
- Integrate scraped data with existing dataset.
  Project Connection: Optionally enrich your dataset with new features.

Day 9 – BeautifulSoup Web Scraping

Task:

- Extract tables, links, titles, or prices from websites.
- Convert scraped data into DataFrame.
- Clean and merge with main dataset.
  Project Connection: Add real-world data for analysis or comparison.

Day 10 – Final Integration

Task:

- Prepare a report or notebook documenting:
  - Dataset description
  - Cleaning steps
  - Key formulas/functions used
  - Visualizations
  - Insights & recommendations
  - Upload the Document with code in .ipynb format with Readme file in GitHub and share the link

Final Outcome:

1. Python Jupyter Notebook (.ipynb) with complete code, cleaned dataset, and visualisations.

2. Documentation (Word/PDF) explaining workflow, insights, and learnings.
3. Upload your final output project to GITHUB with a README file and post your link in the following spreadsheet:

Spreadsheet link:- Coding challenge submission sheet

NOTES:

- You may use any one or multiple datasets from the provided list proived the same dataset should used for all challenges.

- Each day's coding challenge directly contributes to your final project.

- End goal: A complete portfolio-ready Python data analytics project.

Scenario-Based Question Aligned with HackerEarth Codeathon Questions:

Data Analysis Question:

Healthcare Patient Data Analysis:

You are working as a data analyst in a hospital research department. You are tasked with analysing patient health data to identify patterns that can help in preventive healthcare planning.

Dataset: patients.csv-
http://raw.githubusercontent.com/GeethaGunasekaran1/Dataset_rep/refs/heads/main/patients.csv

Columns:

- patient_id: Unique identifier for each patient
- age: Age of the patient
- gender: Gender of the patient
- blood_pressure: Recorded blood pressure value
- cholesterol: Cholesterol level
- diagnosis: Diagnosis category (e.g., Diabetes, Heart Disease, Normal)
- visit_date: Date of hospital visit

Questions:

1. **Patient Overview**
    - How many unique patients are in the dataset?
    - What is the average age of patients?
    - Find the distribution (count) of patients by gender.
2. **Diagnosis Analysis**
    - Count the number of patients per diagnosis category.
    - Find the average cholesterol level for each diagnosis.
    - Identify the top 3 most common diagnoses.
3. **Health Indicators**
    - What is the correlation between age and cholesterol?
    - Find the average blood pressure of patients above age 50.
    - Identify patients (IDs) with both high blood pressure (>140) and high cholesterol (>200).
4. **Time-Based Insights**
    - Group by month to find the total number of visits per month.
    - Find the month with the highest number of patients visiting.

Data Analysis Question :-

Bank Transactions Data Analysis

You are a data analyst at a bank, analyzing customer transaction behavior to detect financial trends.

Dataset: transactions.csv- https://raw.githubusercontent.com/GeethaGunasekaran1/Dataset_rep/refs/heads/main/transactions.csv

Columns:

- transaction_id: Unique identifier
- customer_id: Customer ID

- transaction_date: Date of transaction
- transaction_type: Type (Deposit, Withdrawal, Transfer)
- amount: Transaction amount
- branch: Branch location

Questions:

1. **Transaction Summary**
    - Total number of transactions.
    - Total transaction amount.
    - Average transaction amount.
2. **Customer Analysis**
    - Find the top 5 customers by total transaction amount.
    - Calculate the average transaction amount per customer.
    - Identify customers who made more than 20 transactions.
3. **Branch-Level Analysis**
    - Find the total transaction amount per branch.
    - Identify the top 3 branches by transaction volume.
    - Find the average transaction size for each branch.
4. **Time Analysis**
    - Calculate monthly total transaction amounts.
    - Identify the month with the highest number of withdrawals.

- Find correlation between number of transactions and total transaction amount.

Data Analysis Question

University Admission Data Analysis
You are a data analyst for a university admissions office. The goal is to understand applicant trends and academic performance.

Dataset: admissions.csv-

https://raw.githubusercontent.com/GeethaGunasekaran1/Dataset_rep/refs/heads/main/admissions.csv
Columns:

- student_id: Unique identifier
- name: Student name
- age: Age of applicant
- gender: Gender
- gpa: High school GPA
- program: Program applied for (e.g., Engineering, Arts, Business)
- admission_status: Accepted/Rejected
- application_date: Date of application

Questions:

1. **Applicant Summary**
    - Total number of applicants.
    - Number of accepted vs rejected students.
    - Average GPA of applicants.

2. **Program Analysis**
    - Number of applicants per program.
    - Average GPA per program.
    - Top 3 programs with the highest acceptance rates.
3. **Demographic Insights**
    - Average GPA by gender.
    - Median age of accepted students.
    - Distribution of applicants by age group (e.g., <18, 18–22, >22).
4. **Time Trends**
    - Number of applications per month.
    - Month with highest number of applications.
    - Correlation between GPA and admission_status (binary: 1=Accepted, 0=Rejected).

Data Analysis Question

Social Media Engagement Analysis
You are analyzing engagement metrics for a social media platform to improve content strategies.

Dataset: social_media.csv-

https://raw.githubusercontent.com/GeethaGunasekaran1/Dataset_rep/refs/heads/main/social_media.csv
Columns:

- post_id: Unique identifier for each post
- user_id: ID of the user who posted

- post_type: Type of content (Image, Video, Text, Link)
- likes: Number of likes
- comments: Number of comments
- shares: Number of shares
- post_date: Date of the post

Questions:

1. **Engagement Summary**
    - Total number of posts.
    - Average likes, comments, and shares per post.
    - Post with the maximum likes.
2. **Content Analysis**
    - Average engagement (likes+comments+shares) per post type.
    - Top 2 post types with highest average engagement.
    - Distribution of posts by type.
3. **User Behavior**
    - Identify the top 5 users by total engagement received.
    - Find users who have posted more than 10 times.
    - Calculate average likes for posts created by top 5 users.
4. **Time Analysis**
    - Engagement trend by month.

- Find the best performing month in terms of total engagement.
- Plot a time series of likes over time.

Data Analysis Question

Online Learning Platform Data Analysis
You are analyzing learner data for an online education company.

Dataset: learning.csv- https://raw.githubusercontent.com/GeethaGunasekaran1/Dataset_rep/refs/heads/main/learning.csv
Columns:

- student_id: Unique student identifier
- course_id: Course identifier
- enrollment_date: Date of enrollment
- progress: Percentage of course completed
- score: Final test score (if completed)
- status: Completed / In Progress / Dropped

Questions:

1. **Enrollment Summary**
    - Total number of students enrolled.
    - Number of students per course.
    - Average completion rate per course.
2. **Performance Analysis**
    - Average score of students who completed a course.

- Top 3 courses with the highest average scores.
- Correlation between progress and score.

3. **Dropout Analysis**
    - Percentage of students who dropped courses.
    - Courses with the highest dropout rates.
    - Average progress of students who dropped out.
4. **Time Trends**
    - Enrollments per month.
    - Month with maximum enrollments.
    - Average course completion time (for Completed status).