Technical Assessment for Data Scientist Role

Part 1: Programming and Data Manipulation (Python)

1. Data Cleaning and Preparation:

- Given a CSV file with missing values, outliers, and inconsistent data formats, write a Python script to clean the data. The script should:
    - Handle missing values appropriately.
    - Identify and treat outliers.
    - Standardize data formats (e.g., date formats, categorical variables).

CSV Example:

```
ID,Name,Date_of_Birth,Salary,Department
1,John Doe,1985-10-12,50000,Finance
2,Jane Smith,not_available,62000,Marketing
3,Emily Jones,1990-04-15,70000,Engineering
4,Michael Brown,1975-02-20,45000,HR
,Sarah Davis,1988-08-25,,Sales
6,Peter Pan,1983-07-21,55000,Finance
7,Lily Evans,1985/10/25,1200000,HR
8,Tom Riddle,1982-11-30,31000,Marketing
9,Bruce Wayne,1978-03-10,75000,Finance
10,Clark Kent,1979-12-01,90000,Engineering
11,Diana Prince,1984-05-19,not_applicable,Sales
12,Barry Allen,1977-02-30,1000000,Engineering
13,Arthur Curry,1983-10-10,65000,not_specified
14,Hal Jordan,,72000,HR
15,Victor Stone,1989-11-15,-45000,Marketing
16,Lois Lane,1980-08-20,58000,Finance
17,James Gordon,1975-09-15,52000,HR
18,Selina Kyle,1985-06-07,not_available,Sales
19,Oliver Queen,1982-09-20,1200000,Engineering
20,Roy Harper,1988-07-10,,Marketing
```

2. Data Analysis and Aggregation:

- Using the cleaned dataset, write a Python script to:
    - Calculate the average salary per department.
    - Find the top 3 highest paid employees.
    - Determine the number of employees in each department.

3. API utilization:

- Using the Python code from the previous step, create an API using any web framework. The API should have the following endpoints:
    - An endpoint to get the top N highest-paid employees.
    - An endpoint to get the number of employees in department X.
    - Both X (department) and N (number of employees) should be request parameters.

Part 2: Statistical Analysis

1. Regression Analysis:

- Using a dataset of house prices and features (e.g., size, number of bedrooms, location), perform a linear regression analysis to predict house prices.

Dataset Example:

```
Size, Bedrooms, Location, Price
2000, 3, Urban, 500000
1500, 2, Suburban, 350000
2500, 4, Urban, 750000
1800, 3, Rural, 200000
2200, 3, Suburban, 450000
1600, 2, Urban, 400000
2400, 4, Rural, 300000
1900, 3, Urban, 520000
1700, 2, Suburban, 370000
2300, 4, Urban, 680000
2100, 3, Suburban, 490000
1550, 2, Rural, 220000
2600, 5, Urban, 800000
1750, 3, Suburban, 410000
2000, 3, Rural, 280000
1650, 2, Urban, 390000
2450, 4, Suburban, 610000
1850, 3, Rural, 240000
1700, 2, Urban, 370000
2250, 4, Suburban, 640000
```

Part 3: Data Visualization

1. Dashboard Creation:

- Create a dashboard using a data visualization tool (e.g., Redash, Power BI) to display key metrics from the attached sales dataset. The dashboard should include:
    - Total sales over time.
    - Sales breakdown by product category.
    - Top performing sales regions.

2. Interactive Visualization:

- Using Python (e.g., Plotly, Matplotlib), create an interactive visualization to explore trends in a dataset of your choice (e.g., COVID-19 cases, stock prices).

Instructions for Candidates

1. Submission:

- Provide your solutions in a Python script on a github repo.
- Include comments and explanations for each step of your code.
- For the data visualization tasks, submit screenshots or a link to the interactive dashboard.

2. Evaluation Criteria:

- Accuracy and completeness of solutions.
- Code quality and readability.
- Ability to handle data cleaning and preprocessing.
- Correct application of statistical techniques.
- Effectiveness and clarity of data visualizations.

3. Tools and Libraries:

- Use Python for programming tasks.
- For data visualization, you may use tools like Redash or Power BI.