**CHECKPOINT 1**

|  |  |
| --- | --- |
| **Bộ môn:** Coding | **Năm môn:** Computer Scientist – Advanced |
| **Độ tuổi học viên:** 14 - 18 Tuổi | **Thời lượng:** 60 phút |

# **TRẮC NGHIỆM**

|  |  |
| --- | --- |
| **CÂU HỎI 1:** | Đâu là ví dụ về nguồn dữ liệu phù hợp để phân tích? |
| A. | Wikipedia |
| B. | Kaggle Datasets |
| C. | Dữ liệu được chia sẻ trên Facebook |
| D. | Dữ liệu được chia sẻ trên một bài blog |
| **CÂU HỎI 2:** | Tại sao việc thu thập dữ liệu lại quan trọng trong vòng đời phân tích dữ liệu? |
| A. | Vì thu thập dữ liệu giúp xác định mục tiêu phân tích |
| B. | Vì dữ liệu là nguồn cơ sở để trả lời câu hỏi hoặc giải quyết vấn đề |
| C. | Vì thu thập dữ liệu giúp trực quan hóa tốt hơn |
| D. | Vì thu thập dữ liệu giúp tăng hiệu suất của máy tính |
| **CÂU HỎI 3:** | Câu hỏi nào sau đây yêu cầu phân tích dữ liệu thống kê để trả lời? |
| A. | "Học sinh nào đã đạt điểm cao nhất trong kỳ thi?" |
| B. | "Điểm trung bình của lớp trong kỳ thi là bao nhiêu?" |
| C. | "Ai là người tham dự lớp đầu tiên?" |
| D. | "Môn học nào được yêu thích nhất?" |
| Đáp án đúng là: B | |
| **CÂU HỎI 4:** | Bạn cần phân tích dữ liệu để hiểu rõ hơn về sở thích (thích nhiều – bình thường - không thích) của mọi người, bạn sẽ ưu tiên sử dụng loại dữ liệu nào? |
| A. | Định lượng |
| B. | Định tính |
| C. | Cả hai |
| D. | Không cần sử dụng dữ liệu |
| **CÂU HỎI 5:** | Để tạo một Pandas Series từ một dictionary với khóa là các nhãn và giá trị là các số, sử dụng lệnh: |
| A. | pd.Series([1, 2, 3]) |
| B. | pd.Series({'a': 1, 'b': 2, 'c': 3}) |
| C. | pd.DataFrame({'a': 1, 'b': 2, 'c': 3}) |
| D. | pd.Series(['a', 1, 'b', 2, 'c', 3]) |
| **CÂU HỎI 6:** | Để xoá cột ‘Age’ khỏi DataFrame df, sử dụng câu lệnh: |
| A. | df.remove(‘Age’) |
| B. | df.drop(columns=[‘Age’]) |
| C. | df.delete(‘Age’) |
| D. | df.clear(‘Age’) |
| **CÂU HỎI 7:** | Giả sử bạn có DataFrame df chứa cột ‘Date’. Làm thế nào để lọc các hàng có ‘Date’ từ ngày 1/1/2024 trở đi?  Biết ngày tháng năm có format YYYY-MM-DD. |
| A. | df[df['Date'] >= '2024-01-01'] |
| B. | df[df['Date'] >= pd.datetime('2024-01-01')] |
| C. | df.loc['Date' >= '2024-01-01'] |
| D. | df.query('Date >= "2024-01-01"') |
| **CÂU HỎI 8:** | Bạn có một DataFrame df chứa cột "Department" và "Salary". Làm thế nào để tính trung bình lương cho mỗi phòng ban? |
| A. | df.groupby('Department').avg('Salary') |
| B. | df.groupby('Department')['Salary'].mean() |
| C. | df.groupby('Department').mean('Salary') |
| D. | df.group\_by('Department').mean('Salary') |
| Đáp án đúng là: B | |
| **CÂU HỎI 9:** | Bạn có một DataFrame df chứa các cột "Name", "Age", và "Salary". Dữ liệu có các vấn đề sau:   * Một số giá trị trong cột "Age" bị thiếu (NaN). * Cột "Salary" chứa dữ liệu dưới dạng chuỗi với ký hiệu tiền tệ, ví dụ: "$5000". * Cột "Name" có chứa các khoảng trắng thừa ở đầu và cuối.   Bạn cần làm sạch dữ liệu để chuẩn bị cho phân tích. Hãy chọn các bước làm sạch dữ liệu thích hợp cần thực hiện. |
| A. | 1. Thay thế các giá trị bị thiếu trong cột “Age” bằng giá trị trung bình của cột  2. Chuyển đổi cột “Salary” sang kiểu số bằng cách bỏ ký hiệu tiền tệ  3. Sử dụng phương thức strip() để loại bỏ các khoảng trắng bị thừa trong cột “Name” |
| B. | 1. Thay thế các giá trị bị thiếu trong cột “Age” bằng giá trị trung bình của cột  2. Chuyển đổi cột “Salary” sang kiểu số bằng cách bỏ ký hiệu tiền tệ  3. Loại bỏ cột “Name” ra khỏi DataFrame vì chứa khoảng trắng thừa |
| C. | 1. Thay thế các giá trị bị thiếu trong cột “Age” bằng giá trị trung bình của cột  2. Chuyển đổi cột “Salary” sang kiểu số bằng cách bỏ ký hiệu tiền tệ  3. Loại bỏ cột “Name” ra khỏi DataFrame vì chứa khoảng trắng thừa |
| D. | 1. Thay thế các giá trị bị thiếu trong cột “Age” bằng giá trị trung bình của cột  2. Chuyển đổi cột “Salary” sang kiểu số bằng cách bỏ ký hiệu tiền tệ  3. Loại bỏ cột “Name” ra khỏi DataFrame vì chứa khoảng trắng thừa |
| **CÂU HỎI 10:** | Bạn có một DataFrame df chứa cột "Education Level" với các giá trị "Bachelor", "Master", và "PhD". Bạn muốn thay thế giá trị này bằng các số 1, 2, và 3 tương ứng cho từng cấp độ. Hãy chọn cách sử dụng map() đúng. |
| A. | df['Education Level'] = df.map({'Bachelor': 1, 'Master': 2, 'PhD': 3}) |
| B. | df['Education Level'] = df['Education Level'].map({'Bachelor': 1, 'Master': 2, 'PhD': 3}) |
| C. | df['Education Level'] = df['Education Level'].apply(lambda x: 1 if x == 'Bachelor' else 2 if x == 'Master' else 3) |
| D. | df['Education Level'] = df['Education Level'].replace({'Bachelor': 1, 'Master': 2, 'PhD': 3}) |

# **THỰC HÀNH**

|  |  |
| --- | --- |
| **ĐỀ BÀI:** | Sử dụng dữ liệu [“car-sales-data.csv”](https://drive.google.com/file/d/1LpaDRp6jeqYyfkey_QGAo3hWLweMtt3t/view?usp=sharing) của một công ty bán xe cũ,  Hãy phân tích dữ liệu để trả lời cho các câu hỏi sau:   1. Chiếc xe nào được bán với giá đắt nhất? 2. Hãng xe nào có chiếc xe với giá trên mỗi kilomet đã đi được rẻ nhất? 3. Trung bình giá xe của các hãng xe là bao nhiêu? Hãng xe nào có trung bình giá bán cao nhất? |
| **HƯỚNG DẪN:** | Giải nghĩa các cột trong file dữ liệu:   * Make: hãng xe * Colour: màu xe * Odometer: quãng đường xe đã đi được (đơn vị kilomet) * Doors: số cửa * Price: giá xe (đơn vị $)   Lưu ý: Áp dụng các phương pháp làm sạch và xử lý dữ liệu hợp lý trước khi phân tích dữ liệu |