**TRƯỜNG ĐẠI HỌC SƯ PHẠM HÀ NỘI**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO BÀI TẬP LỚN**

**Học phần: COMP358 – Big Data**

**XÂY DỰNG HỆ THỐNG PHÂN TÍCH DỮ LIỆU BẰNG HADOOP – PHÂN TÍCH SỨC KHỎE TÂM THẦN**

**Giảng viên: Thầy Giang Thành Trung**

**MỤC LỤC**

DANH SÁCH THÀNH VIÊN1

GIỚI THIỆU VỀ BÀI TOÁN2

TỔNG QUAN CÔNG NGHỆ SỬ DỤNG 3

THU THẬP VÀ MÔ PHÒNG DỮ LIỆU4

LƯU TRỮ DỮ LIỆU TRÊN HDFS5

THIẾT KẾ TRUY VẤN VỚI HIVE6

TRỰC QUAN HÓA DỮ LIỆU7

PHÂN TÍCH VÀ DỰ ĐOÁN XU HƯỚNG8

1. **DANH SÁCH THÀNH VIÊN**

|  |  |  |
| --- | --- | --- |
| **STT** | **Họ và tên** | **Mã sinh viên** |
| 1 | Nguyễn Hoàng Tùng | 735105112 |
| 2 | Nguyễn Huy Hải Ninh | 735105084 |
| 3 | Trần Gia Lộc | 735105060 |
| 4 | Tạ Minh Quang | 735105088 |

1. **GIỚI THIỆU BÀI TOÁN**

* Sức khỏe tâm thần ngày càng trở thành một vấn đề quan trọng trong xã hội hiện đại, đặc biệt là đối với nhóm tuổi thanh niên và trung niên (18 - 60 tuổi). Các vấn đề về sức khỏe tâm thần như lo âu, trầm cảm, căng thẳng và stress có thể ảnh hưởng nghiêm trọng đến chất lượng cuộc sống, khả năng làm việc, cũng như sự phát triển của cá nhân trong xã hội. Trong bối cảnh này, việc phân tích và hiểu rõ các yếu tố ảnh hưởng đến sức khỏe tâm thần trở nên cấp thiết.
* Bài toán này tập trung vào việc **phân tích dữ liệu sức khỏe tâm thần** của nhóm tuổi thanh niên và trung niên (18 - 60 tuổi) với mục tiêu nhận diện các yếu tố tác động đến sức khỏe tâm thần, từ đó đưa ra những dự báo và giải pháp can thiệp kịp thời.

1. **TỔNG QUAN CÔNG NGHỆ SỬ DỤNG**
2. **Apache Hadoop**

* Hadoop cung cấp nền tảng lưu trữ phân tán với **Hadoop Distributed File System (HDFS)**, giúp lưu trữ và quản lý khối lượng dữ liệu lớn.

1. **Apache Hive**

* Hive cho phép truy vấn bằng HiveQL, hỗ trợ lưu trữ phân tán, xử lý song song và tối ưu khi làm việc với dữ liệu lớn.

1. **Python**

* Tạo dữ liệu giả lập
* Ghi file CSV
* Chuyển đổi file CSV sang Parquet

1. **Power BI**

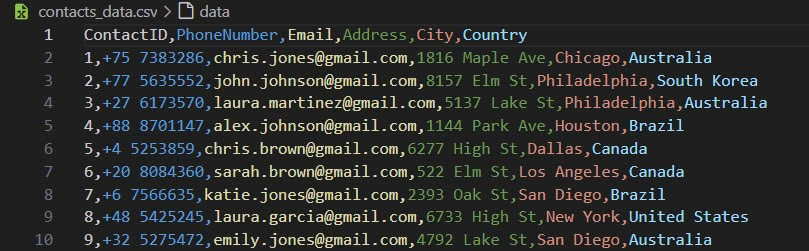
* Trực quan hóa kết quả phân tích dữ liệu dưới dạng biểu đồ,...

1. **THU THẬP VÀ MÔ PHỎNG DỮ LIỆU**

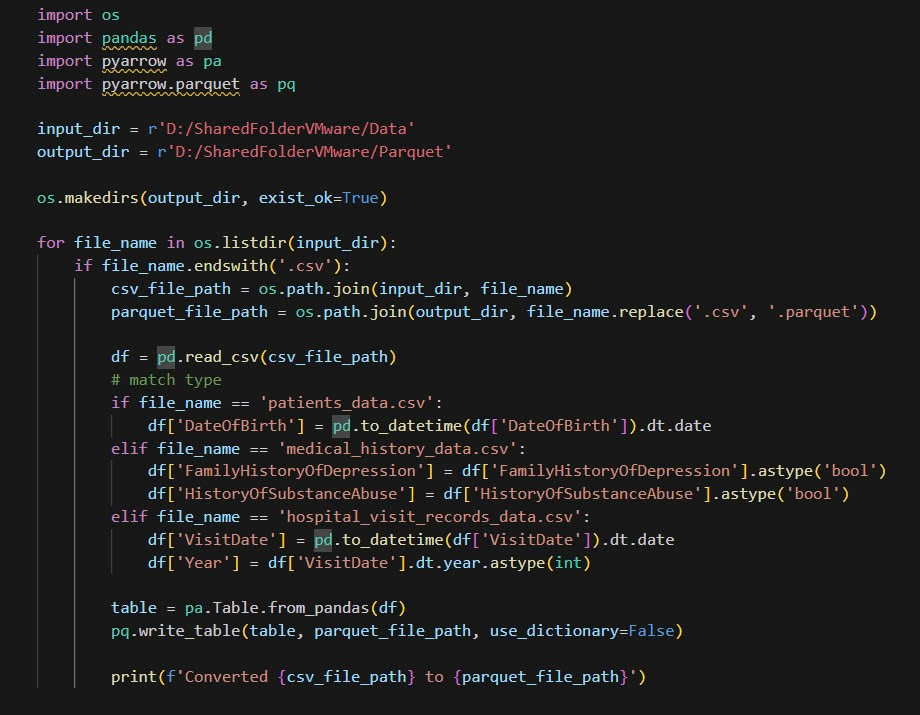
* Tham khảo một số nguồn mô hình dữ liệu trên mạng.
* Tạo dữ liệu giả lập bằng thư viện Random và List tự tạo của Python.



* Sử dụng thư viện Random với các hàm để tạo ra các giá trị ngẫu nhiên
* Hàm **genernate\_contact\_row** sinh ra dữ liệu cho từng dòng
  + row\_id
  + số điện thoại
  + …….
* Hàm **generate\_contact\_data** tạo dữ liệu ghi ra file CSV
* Đoạn mã **if \_\_name\_\_ == “ \_\_main\_\_”** đảm bảo file này sẽ được chạy trực tiếp, không bị chạy trùng lặp.
* Kết quả:



* Sử dụng định dạng Parquet để tối ưu lưu trữ và truy vấn trên Hive
* Dữ liệu lưu dưới dạng cột thay vì hàng
* Chỉ cần tìm kiếm trên cột.
* Tối ưu hóa dung lượng dữ liệu (Dung lượng file Parquet nhỏ hơn dung lượng file CSV).
* Các bước chuyển đổi:



+ Sử dụng thư viện **os, pandas, pyarrow.**

**+** Duyệt các tên file trong thư mục Input

++ Khai báo địa chỉ của file đầu vào (.csv).

` ++ Khái báo địa chỉ của file đầu ra (.parquet).

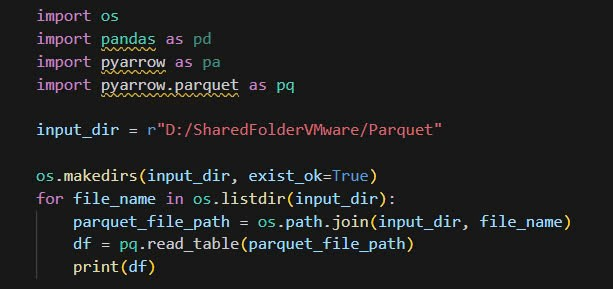
+ Khởi tạo một **data frame** thông qua hàm **read\_csv**

**+** Xử lí một số trường mà Hive không thể tự xác định kiểu dữ liệu (Datetime, bool,….).

+ Khởi tạo biến table có giá trị được chuyển đổi từ **data frame** của **pandas** sang **table** của **pyarrow** để phù hợp với định dạng **parquet**.

+ Dùng hàm **write\_table** của **module parquet** trong **pyarrow** để viết ra file **parquet.**

* Kiểm tra dữ liệu:



+ Sử dụng hàm **read\_table** của **module parquet** trong **pyarrow** để gán giá trị cho biến, rồi in ra biến và hiển thị kết quả.

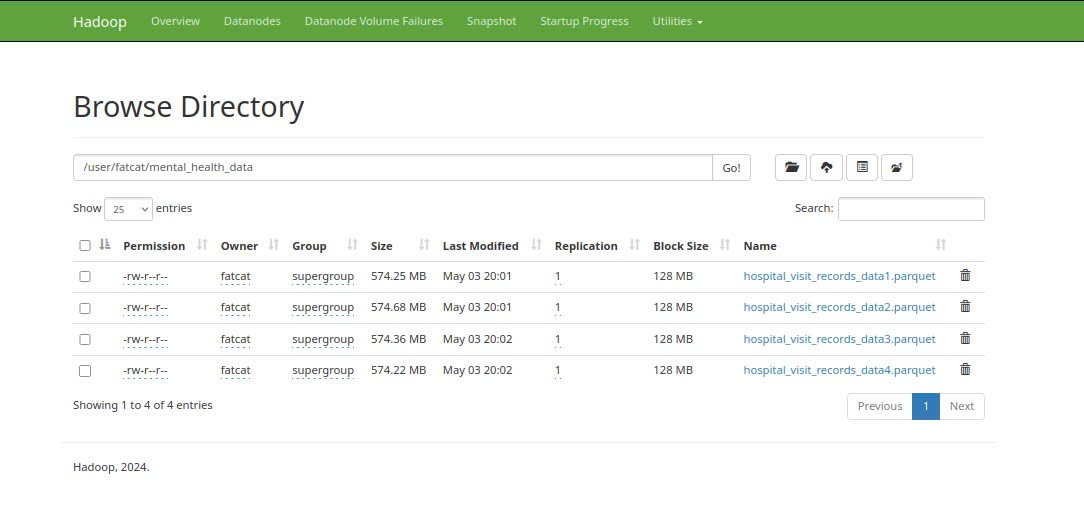
+ Kết quả:



1. **LƯU TRỮ DỮ LIỆU TRÊN HDFS Và Hive**

* Sử dụng lệnh Hadoop để tải dữ liệu Parquet lên HDFS.





* Tạo bảng Hive cho cơ sở dữ liệu:



* Thêm dữ liệu vào table từ file Parquet lưu trên hệ thống HDFS:

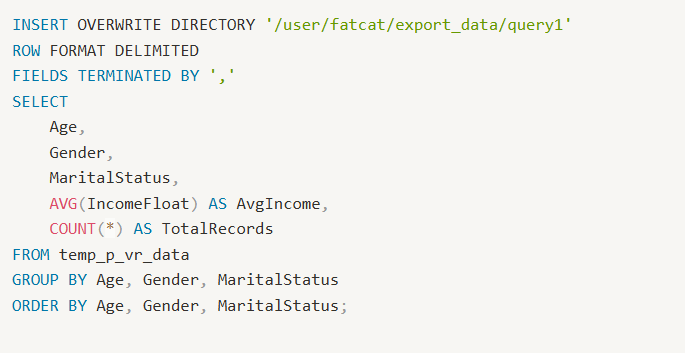


1. **THIẾT KẾ TRUY VẤN VỚI HIVE**

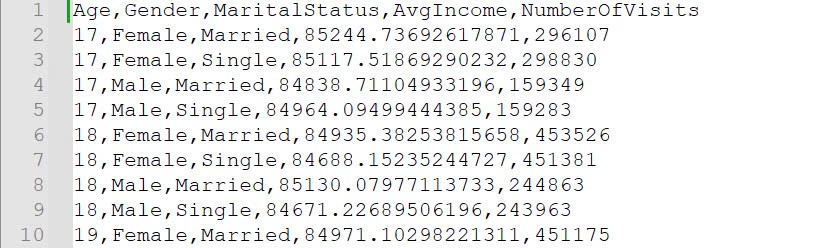
* Sử dụng **INSERT OVERWRITE DIRECTORY** để ghi dữ liệu vào thư mục **'/user/fatcat/export\_data/.....'.**
* Dữ liệu được xuất theo định dạng **ROW FORMAT DELIMITED** với các trường được phân tách bằng dấu phẩy (,).

1. **Thống kê số bệnh nhân theo tuổi, giới tính, tình trạng hôn nhân và thu nhập**

* **Xu hướng phân tích:**
  + **So sánh thu nhập trung bình theo độ tuổi, giới tính, và tình trạng hôn nhân:**
    - Từ đó sẽ thấy được khác biệt về thu nhập trung bình giữa các nhóm.
  + **Sự thay đổi giữa các nhóm:**
    - Từ đó, có thể so sánh sự thay đổi giữa các nhóm độ tuổi, giới tính và tình trạng hôn nhân về thu nhập trung bình.

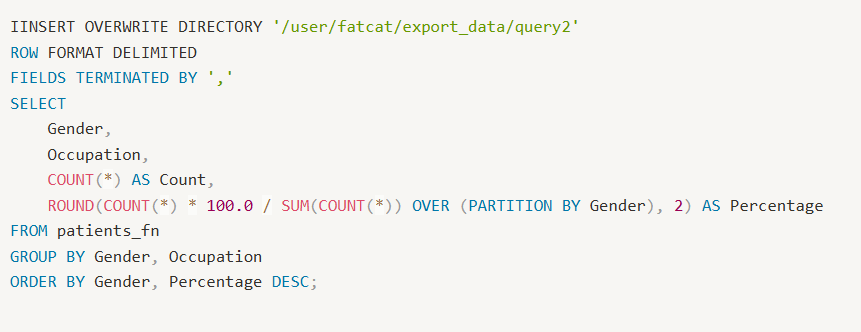
****

* Chọn dữ liệu:
  + Age: tuổi
  + Gender: giới tính
  + MaritalStatus: tình trạng hôn nhân
  + AvgIncome: thu nhập trung bình
  + TotalRecords: tổng số bản ghi trong mỗi nhóm
* Kết quả:

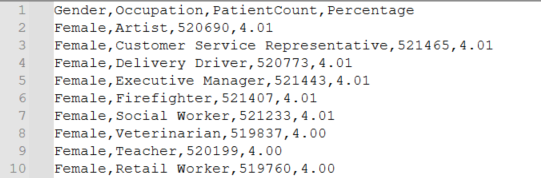


1. **Thống kê số bệnh nhân theo giới tính, nghề nghiệp và tỷ lệ phần tram**

* **Xu hướng phân tích:**
  + **Phân tích tỷ lệ phần trăm nghề nghiệp theo giới tính:**
    - Từ đó thấy được các nghề nghiệp nào chiếm tỷ lệ cao hơn ở giới tính nào, từ đó có thể đưa ra nhận định về sự phân hóa giới tính trong từng ngành nghê.
  + **So sánh giữa các nghề nghiệp trong mỗi giới tính:**
    - Giúp nhận diện xu hướng nghề nghiệp phổ biến nhất cho từng giới tính, từ đó kết luận về sự phân bố nghề nghiệp theo giới tính.

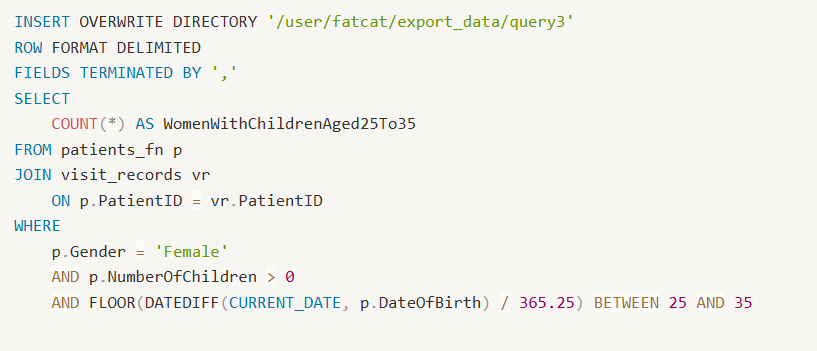
****

* Chọn dữ liệu:
  + Gender: Giới tính
  + Occupation: Nghề nghiệp
  + Count: Đếm số lượng bản ghi cho từng nhóm Gender và Occupation
  + Percentage: Tính phần tram số lượng bản ghi của từng Occupation trong tổng số bản ghi của từng Gender.
* Kết quả:



1. **Thống kê số bệnh nhân phụ nữ có con độ tuổi từ 25 đến 35**

* **Xu hướng phân tích:**
  + **Phân tích tổng số phụ nữ có con trong độ tuổi từ 25 đến 35:**
    - Từ đó hiểu rõ hơn về nhu cầu, hành vi, và đặc điểm của nhóm này.
  + **Phân tích xu hướng sinh con theo độ tuổi:**
    - Từ đó có cái nhìn về xu hướng sinh đẻ ở các nhóm độ tuổi khác nhau, đưa ra kết luận về độ tuổi nào là giai đoạn phổ biến để sinh con.

****

* Chọn dữ liệu:
  + Đếm số lượng phụ nữ có con và nằm trong độ tuổi từ 25 đến 35
* Kết quả:



1. **Số lượt khám theo từng năm, tình trạng sức khỏe chung**

* **Xu hướng phân tích:**
  + **Phân tích xu hướng thay đổi tình trạng sức khỏe theo năm:**
    - Giúp nhận ra xu hướng sức khỏe của cộng đồng trong thời gian dài và xác định xem tình trạng sức khỏe có cải thiện hay xấu đi theo từng năm hay không.
  + **So sánh tỷ lệ bệnh nhân trong các tình trạng sức khỏe:**
    - Từ đó đánh giá mức độ cải thiện hoặc suy giảm trong sức khỏe của bệnh nhân.

****

* Chọn dữ liệu:
  + Year: năm
  + GeneralHealthStatus: Tình trạng sức khỏe chung (Good, Fair, Poor,…)
  + PatientCount: Đếm số lượng bệnh nhân cho từng nhóm.
* Kết quả:

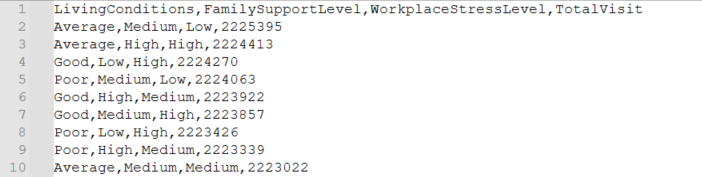


1. **Thống kê điều kiện sống, mức độ hỗ trợ gia đình và mức độ căng thẳng tại nơi làm việc**

* **Xu hướng phân tích:**
  + **Phân tích sự kết hợp giữa các yếu tố sống, hỗ trợ gia đình và căng thẳng công việc:**
    - Giúp nhận biết nhóm nào có các yếu tố tiêu cực và nhóm nào có sự hỗ trợ tốt từ gia đình, từ đó đưa ra kết luận về sự tác động của các yếu tố xã hội.
  + **So sánh tổng số lượng thăm khám (TotalVisit) theo các yếu tố**:
    - Việc này sẽ giúp nhận ra mối liên hệ giữa mức độ thăm khám với tình trạng sống và công việc, và từ đó đưa ra kết luận về các yếu tố xã hội ảnh hưởng đến sức khỏe cộng đồng.

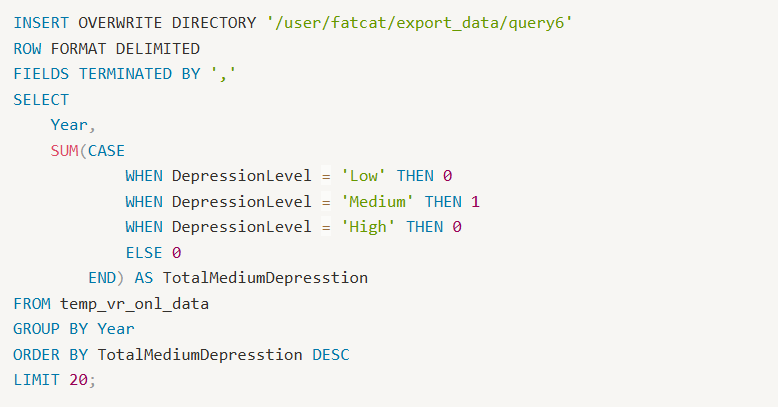
****

* Chọn dữ liệu:
  + LivingConditions: Điều kiện sống
  + FamilySupportLevel: Mức độ hỗ trợ từ gia đình
  + WorkplaceStressLevel: Mức độ căng thẳng tại nơi làm việc
  + TotalVisit: Đếm tổng số bản ghi cho từng nhóm
* Kết quả:

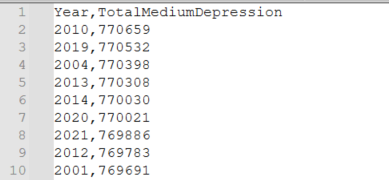


1. **Thống kê năm có số lượng bệnh nhân trầm cảm mức trung bình nhiều nhất**

* **Xu hướng phân tích:**
  + **Phân tích sự thay đổi tổng số trường hợp trầm cảm mức độ trung bình qua các năm**:
    - Giúp nhận diện xu hướng tăng giảm của các trường hợp trầm cảm trung bình theo thời gian, từ đó đưa ra kết luận về sự biến động tình trạng trầm cảm trong cộng đồng.
  + **So sánh giữa các năm về số lượng trường hợp trầm cảm trung bình**:
    - Điều này giúp phân tích các năm có sự gia tăng hoặc giảm sút đáng chú ý trong mức độ trầm cảm trung bình.

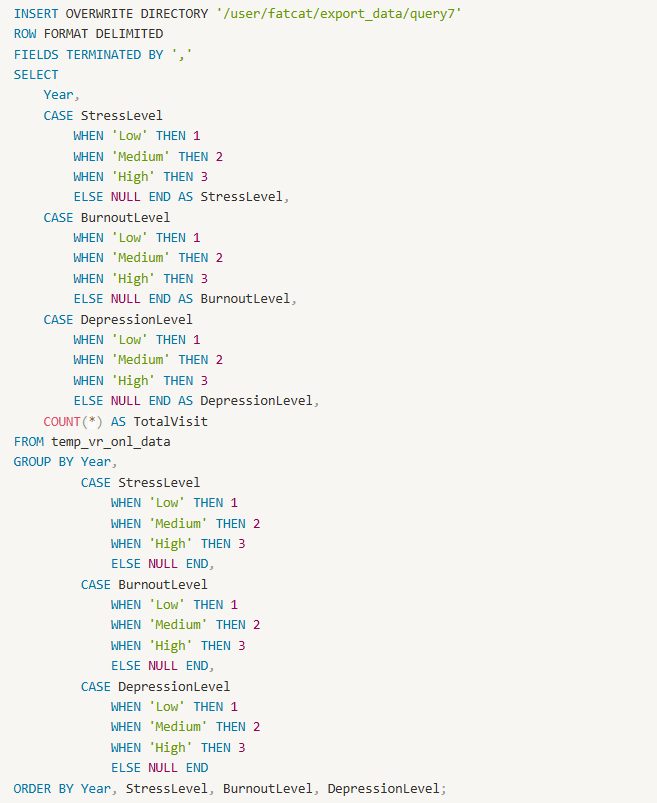


* Chọn dữ liệu:
  + Year: năm
  + TotalMediumDepression: số lượng trường hợp mắc bệnh trầm cảm ở mức trung bình:
    - Nếu ở mức “Low” hoặc “High” thì giá trị được quy đổi là 0.
    - Nếu ở mức “Medium” thì giá trị được quy là 1.
    - Chỉ lọc ra giá trị “Medium”.
* Kết quả:

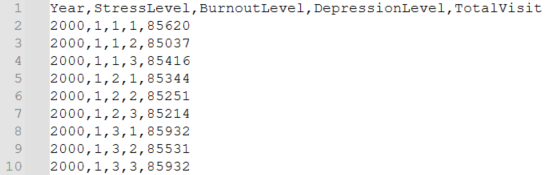


1. **Thống kê số bệnh nhân theo năm và mức độ căng thẳng, kiệt sức, trầm cảm**

* **Xu hướng phân tích:**
  + **Phân tích sự thay đổi mức độ căng thẳng, kiệt sức và trầm cảm qua các năm**:
    - Điều này giúp nhận diện các năm có sự gia tăng hoặc giảm sút đáng chú ý về các vấn đề tâm lý, từ đó đánh giá tác động của các yếu tố xã hội, môi trường hay chính sách đối với sức khỏe tâm lý cộng đồng.
  + **So sánh mức độ căng thẳng, kiệt sức và trầm cảm trong mỗi năm**:
    - Giúp nhận diện mối quan hệ giữa các yếu tố này và xác định yếu tố nào có ảnh hưởng mạnh mẽ nhất trong từng năm. Việc này sẽ giúp hiểu rõ sự phân bổ và mức độ nghiêm trọng của từng vấn đề tâm lý trong các năm qua.

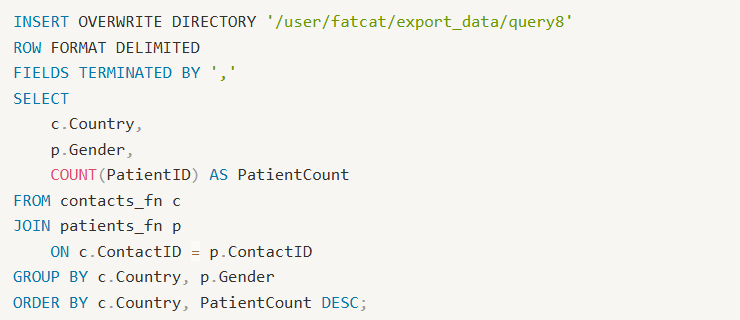
****

* Chọn dữ liệu:
  + Year: năm
  + StressLevel: Chuyển các giá trị của StressLevel thành số:
    - “Low” : 1
    - “Medium”: 2
    - “High”: 3
    - Các giá trị khác thành null
  + BurnoutLevel: Chuyển các giá trị của BurnoutLevel thành số tương tự StressLevel.
  + DepressionLevel: Chuyển các giá trị của DepressionLevel thành số tương tự StressLevel.
  + TotalVisits: Đếm tổng số bản ghi cho từng nhóm trên.
* Kết quả:

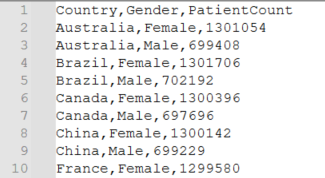


1. **Phân tích số bệnh nhân theo một số nước và giới tính**

* **Xu hướng phân tích:**
  + **Phân tích phân bố bệnh nhân theo quốc gia**:
    - Xác định quốc gia nào có số lượng bệnh nhân cao nhất. Điều này giúp nhận diện các quốc gia có tỷ lệ bệnh nhân cao và đánh giá tác động của các yếu tố quốc gia đến sức khỏe cộng đồng.
  + **So sánh sự phân bố giới tính theo quốc gia**:
    - Nhận diện các quốc gia có tỷ lệ bệnh nhân nam hoặc nữ chiếm ưu thế. Điều này giúp hiểu rõ sự phân hóa giới tính của bệnh nhân và các yếu tố xã hội có thể tác động đến sức khỏe giữa nam và nữ tại các quốc gia khác nhau.

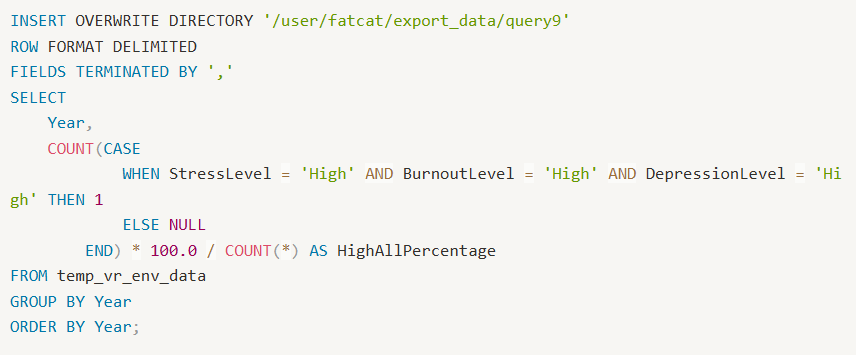
****

* Chọn dữ liệu:
  + c.Country: quốc gia
  + p.Gender: Giới tính
  + PatientCount: Đếm số lượng bệnh nhân (PatientID) cho từng nhóm Country và Gender
* Kết quả:

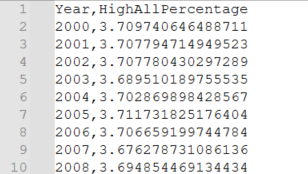


1. **Thống kê tỉ lệ số bệnh nhân có ba mức độ căng thẳng, kiệt sức, trầm cảm cao nhất**

* **Xu hướng phân tích:**
  + **Phân tích tỷ lệ người có mức độ căng thẳng, kiệt sức và trầm cảm cao theo năm**:
    - Nhận diện xu hướng gia tăng hoặc giảm sút của các vấn đề này. Điều này giúp bạn hiểu rõ hơn về sự thay đổi tình trạng sức khỏe tâm lý trong cộng đồng qua thời gian.
  + **So sánh sự thay đổi tỷ lệ giữa các năm**:
    - Xác định các năm có sự gia tăng hoặc giảm sút đáng chú ý trong tỷ lệ người gặp phải căng thẳng, kiệt sức và trầm cảm cao. Việc này sẽ giúp phân tích các yếu tố có thể tác động đến sự thay đổi này, như biến động xã hội, chính sách y tế, hay các yếu tố môi trường.

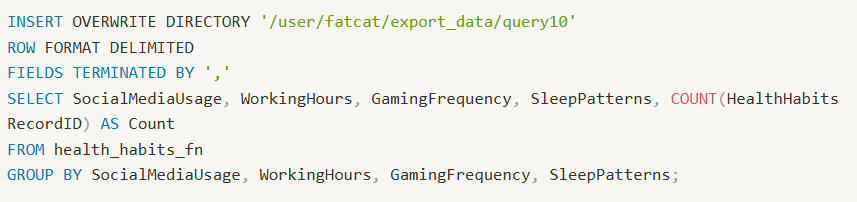
****

* Chọn dữ liệu:
  + Year: năm
  + HighAllPercentage: Tính phần tram số lượng bản ghi có cả ba mức độ StressLevel, BurnoutLevel, DepressionLevel đều là “High” so với tổng số bản ghi trong từng năm:
    - Nếu cả ba mức độ là “High”, giá trị được quy đổi là 1.
    - Ngược lại là null.
* Kết quả:

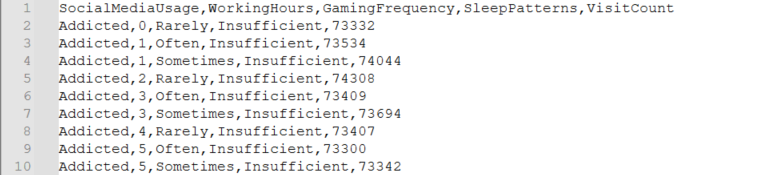


1. **Phân tích xem một số các yếu tố ảnh hưởng đến thói quen ngủ**

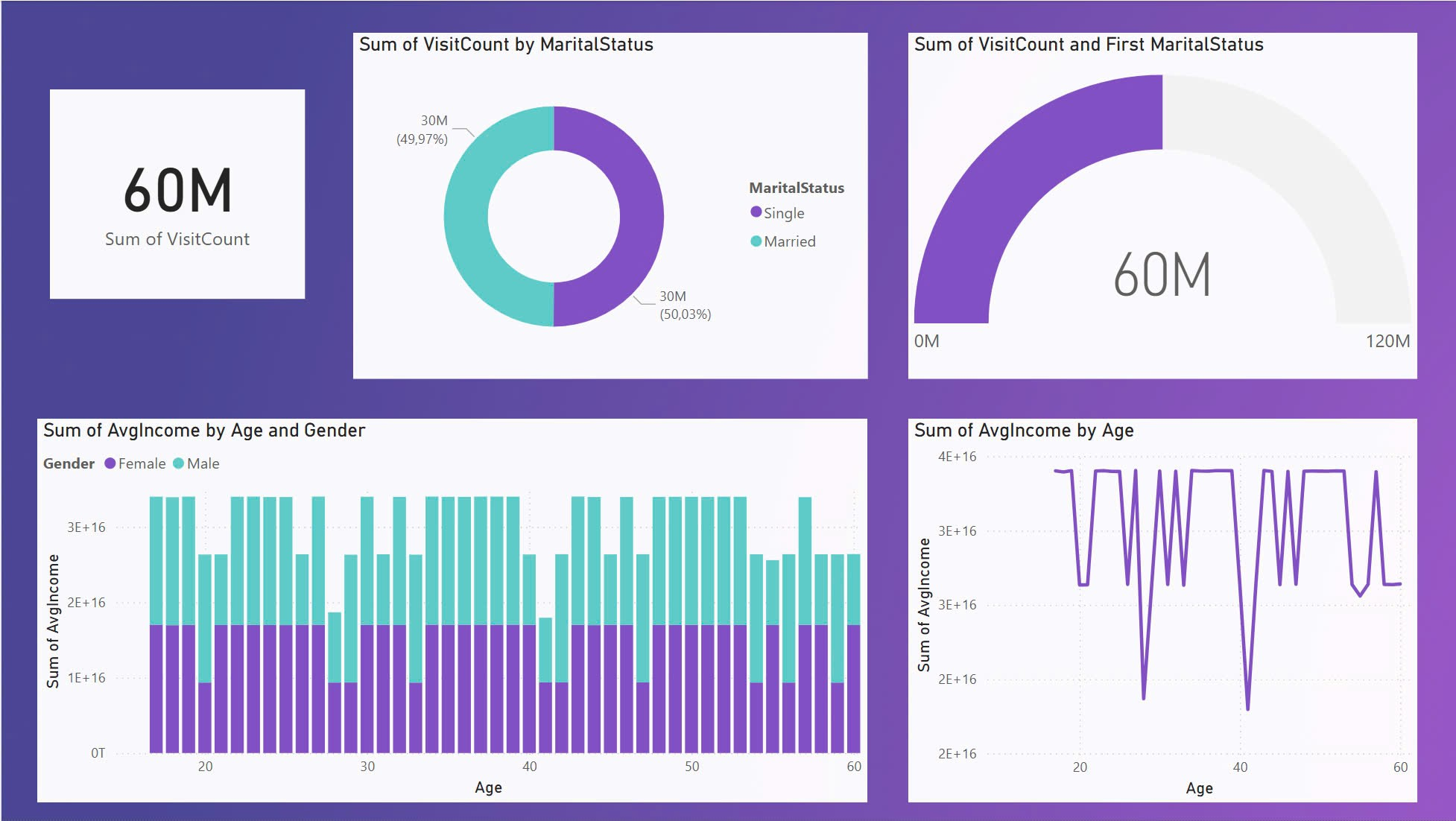
* **Xu hướng phân tích:**
  + **Phân tích sự phân bố thói quen sức khỏe theo các yếu tố**:
    - Nhận diện các yếu tố nào có ảnh hưởng lớn nhất đến tình trạng sức khỏe. Việc này sẽ giúp xác định những nhóm có thói quen sức khỏe tốt hay xấu và từ đó đưa ra các biện pháp cải thiện.
  + **So sánh giữa các nhóm thói quen sức khỏe**:
    - Điều này giúp nhận diện những yếu tố có ảnh hưởng mạnh mẽ đến sức khỏe, như thời gian ngủ hoặc thời gian dành cho hoạt động giải trí.



* Chọn dữ liệu:
  + SocialMediaUsage: Mức độ sử dụng mạng xã hội
  + WorkingHours: Số giờ làm việc
  + GamingFrequency: Tần suất chơi game
  + SleepPatterns: Thói quen ngủ
  + Count: Đếm số lượng bản ghi (HealthHabitsRecordID) cho từng nhóm.
* Kết quả:



1. **TRỰC QUAN HÓA DỮ LIỆU**
2. **Phân tích dữ liệu người dùng theo tình trạng hôn nhân, giới tính và độ tuổi**

****

1. **Mục tiêu phân tích**

* Đánh giá số lượt đến thăm khám của người dùng thông qua chỉ số VisitCount.
* Phân tích mức thu nhập trung bình (AvgIncome) theo độ tuổi và giới tính.
* Phân loại hành vi người dùng dựa trên tình trạng hôn nhân (độc thân – kết hôn).

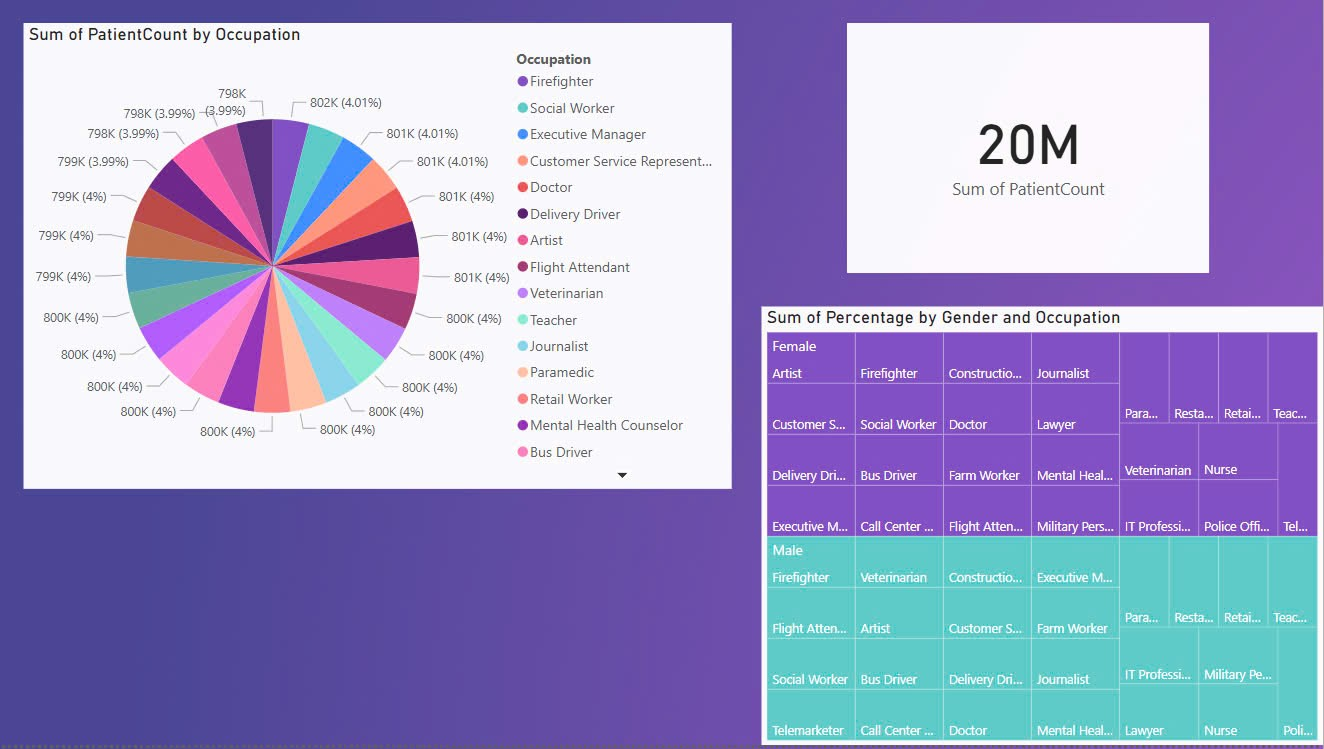
1. **Tổng quan dữ liệu**

* Tổng số lượt VisitCount: 60 triệu.
* Dữ liệu được trình bày trực quan thông qua các biểu đồ: doughnut, gauge, biểu đồ côt, và biểu đồ đường.

1. **Phân tích chi tiết**

* **Tình trạng hôn nhân**
  + Tỷ lệ VisitCount theo tình trạng hôn nhân:
    - Single: ~30 triệu (49,97%).
    - Married: ~30 triệu (50,03%).
  + Nhận định: Hai nhóm người dùng có tỷ lệ đến thăm khám gần như bằng nhau. Điều này cho thấy tình trạng hôn nhân không tạo ra sự chênh lệch lớn trong hành vi truy câp.
* **Tổng lượt đến thăm khám và phân phối theo tình trạng kết hôn đầu tiên**
  + Gauge chart hiển thị 60 triệu lượt, tương ứng với 50% giới hạn tối đa (120 triệu).
  + Nhận định: Nguồn dữ liệu hiện tại mới chỉ phản ánh một nửa tiềm năng phân tích – còn nhiều dư địa để mở rộng và đào sâu dữ liệu.
* **Phân tích thu nhập theo giới tính và độ tuổi**
  + Biểu đồ cột chồng (stacked column chart) thể hiện:
    - Nam giới (Màu xanh) có mức thu nhập trung bình cao hơn nữ trong đa số độ tuổi.
    - Đặc biệt trong độ tuổi từ 25 – 45, nam giới chiếm tỷ trọng thu nhập lớn hơn rõ rệt.
    - Nữ giới (Màu tím) có xu hướng giữ mức thu nhập ổn định nhưng thấp hơn.
  + Nhận định: Có sự khác biệt đáng kể về thu nhập theo giới.
* **Biến động thu nhập theo độ tuổi**
  + Biểu đồ đường cho thấy:
    - Mức thu nhập trung bình biến động mạnh theo từng nhóm tuổi.
    - Một số điểm rơi bất thường xảy ra trong khoảng tuổi từ 30 – 40.
  + Nhận định: Các điểm rơi này có thể do các biến động ngoài dự đoán như thay đổi nghề nghiệp hoặc tình trạng tài chính cá nhân trong giai đoạn này.

1. **Phân tích dữ liệu người dùng theo nghề nghiệp và giới tính**

****

1. **Mục tiêu phân tích**

* Phân tích số lượng bệnh nhân (PatientCount) theo các nghề nghiệp và giới tính.
* Đánh giá tỷ lệ phần trăm bệnh nhân trong các nhóm nghề nghiệp theo giới tính.

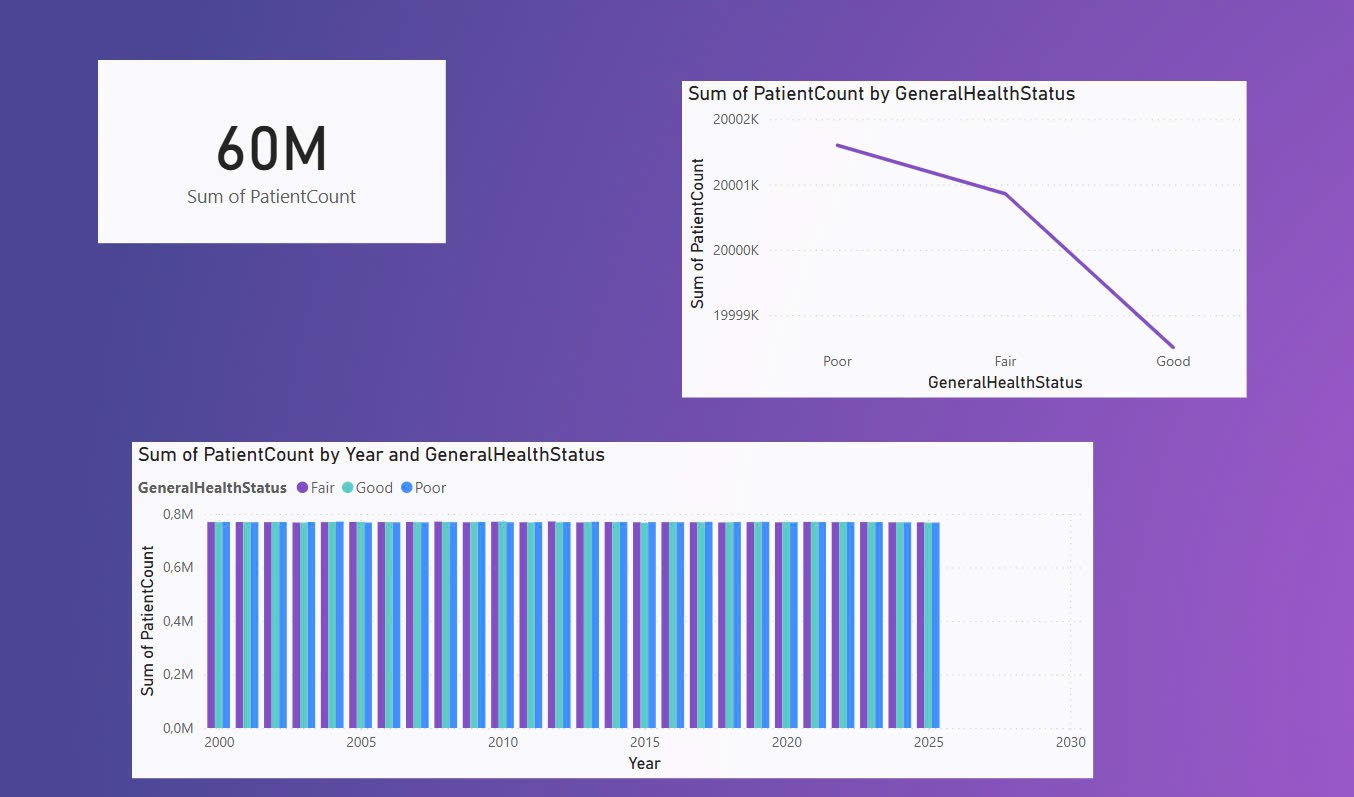
1. **Tổng quan dữ liệu**

* Tổng số bệnh nhân (PatientCount): 20 triệu người.
* Các biểu đồ: Biểu đồ bánh (Pie Chart) và Biểu đồ tỷ lệ phần tram theo nghề nghiệp và giới tính.

1. **Phân tích chi tiết**

* **Phân bố bệnh nhân theo nghề nghiệp**
  + Biểu đồ bánh (Pie Chart):
    - Các nghề nghiệp như Lính cứu hỏa, Quản lý điều hành, Nhân viên dịch vụ khách hàng, …. chiếm tỷ lệ 4% mỗi nhóm.
    - Sự phân bố của bệnh nhân tương đối đồng đều, với các ngành nghề chủ yếu như Lính cứu hỏa và Quản lý điều hành chiếm tỷ lệ lớn.
  + Nhận định: Các nhóm nghề nghiệp có sự phân bố tương đối đồng đều. Tuy nhiên, một số nhóm như Nhân viên bán lẻ và Giáo viên có tỷ lệ thấp hơn, cần có chiến lược chăm sóc sức khỏe phù hợp với các nhóm này.
* **Tỷ lệ phần trăm bệnh nhân theo giới tính và nghề nghiệp**
  + Biểu đồ tỷ lệ phần trăm
    - **Nữ giới** chiếm ưu thế trong các nhóm nghề nghiệp như **Nghệ sĩ, Lính cứu hỏa,** và **Công tác xã hội**.
    - **Nam giới** chiếm ưu thế trong các nghề như **Bác sĩ thú y, Tiếp viên hàng không, Nhà báo, và Giám đốc điều hành.**
    - Các nhóm nghề nghiệp có sự phân bổ khá đồng đều giữa nam và nữ, nhưng vẫn có một số ngành nghề có sự phân hóa rõ rệt về giới tính.
  + Nhận định: Các ngành nghề như **Nghệ sĩ** và **Lính cứu hỏa** có tỷ lệ nữ giới cao, trong khi các nghề như **Bác sĩ thú y** và **Giám đốc điều hành** lại thiên về nam giới. Điều này có thể phản ánh xu hướng nghề nghiệp và sức khỏe liên quan đến giới tính.

1. **Phân tích dữ liệu liệu sức khỏe của bệnh nhân theo trạng thái sức khỏe và thời gian**

****

1. **Mục tiêu phân tích**

* Phân tích số lượng bệnh nhân (PatientCount) theo **trạng thái sức khỏe** tổng thể (GeneralHealthStatus) trong các năm khác nhau.
* Xác định mối quan hệ giữa **trạng thái sức khỏe** (Tốt, Trung bình, Kém) và **tổng số bệnh nhân** theo thời gian.
* Cung cấp cái nhìn tổng thể về **biến động tình trạng sức khỏe** trong suốt giai đoạn từ năm 2000 đến năm 2030.

1. **Tổng quan dữ liệu**

* **Tổng số bệnh nhân (PatientCount):** 60 triệu người.
* **Dữ liệu phân tích:** Bao gồm ba trạng thái sức khỏe **Tốt, Trung bình,** và **Kém**, được trình bày qua các biểu đồ:
  + **Biểu đồ đường (Line Chart)** cho thấy sự thay đổi của **PatientCount** theo **trạng thái sức khỏe.**
  + **Biểu đồ cột theo năm (Bar Chart)** phân bổ số lượng bệnh nhân theo từng năm.

1. **Phân tích chi tiết**

* **Sự thay đổi của bệnh nhân theo trạng thái sức khỏe**
  + Biểu đồ đường (Line Chart):
    - **Trạng thái sức khỏe Kém (Poor)** có xu hướng giảm mạnh so với các trạng thái khác.
    - **Trạng thái sức khỏe Trung bình (Fair)** có sự ổn định và duy trì tỷ lệ **bệnh nhân cao,** nằm giữa nhóm "Tốt" và "Kém".
    - **Trạng thái sức khỏe Tốt (Good)** có xu hướng đạt đỉnh cao nhất, nhưng có sự giảm dần trong vài năm gần đây.
  + Nhận định: **Trạng thái sức khỏe Tốt** đang giảm dần, có thể là do cải thiện chăm sóc sức khỏe hoặc thay đổi lối sống. **Trạng thái Kém** đang giảm, cho thấy sự cải thiện trong việc chăm sóc sức khỏe cộng đồng.
* **Phân bổ bệnh nhân theo năm và trạng thái sức khỏe**
  + Biểu đồ cột theo năm (Bar Chart):
    - **Số bệnh nhân** trong các năm từ 2000 đến 2030 có sự phân bổ khá đều giữa các **trạng thái sức khỏe,** tuy nhiên tỷ lệ bệnh nhân **Tốt** có xu hướng giảm dần.
    - **Trạng thái Trung bình** và **Kém** có sự ổn định hơn, không có sự thay đổi đột biến giữa các năm.
  + Nhận định: **Sự ổn định của bệnh nhân** trong các trạng thái sức khỏe cho thấy các chương trình chăm sóc sức khỏe có thể chưa tạo ra tác động lớn đến việc thay đổi trạng thái sức khỏe trong thời gian dài.

1. **Phân tích dữ liệu về tình trạng sức khỏe và các yếu tố ảnh hưởng**

****

1. **Mục tiêu phân tích**

* Phân tích số lượng **lượt đến thăm (TotalVisit)** theo các yếu tố: **Điều kiện sống, Mức độ hỗ trợ gia đình,** và **Mức độ căng thẳng tại nơi làm việc.**

1. **Tổng quan dữ liệu**

* **Tổng số lượt đến khám (TotalVisit):** 60 triệu lượt.
* Dữ liệu được phân tích qua các biểu đồ:
  + **Biểu đồ thanh (Bar Chart)** hiển thị sự phân bổ **lượt đến thăm khám theo mức độ hỗ trợ gia đình.**
  + **Biểu đồ cột (Column Chart)** cho thấy mối quan hệ giữa **điều kiện sống** và **mức độ hỗ trợ gia đình.**
  + **Biểu đồ tròn (Pie Chart)** cho thấy tỷ lệ **lượt đến thăm khám theo mức độ căng thẳng tại nơi làm việc.**

1. **Phân tích chi tiết**

* **Phân bổ lượt đến thăm khám theo điều kiện sống và mức độ hỗ trợ gia đình**
  + Biểu đồ cột (Column Chart):
    - **Điều kiện sống tốt (Good)** có tỷ lệ lượt truy cập cao nhất, với **mức độ hỗ trợ gia đình cao** chiếm ưu thế.
    - **Điều kiện sống trung bình (Average)** và **kém (Poor)** cũng có lượt truy cập cao, nhưng tỷ lệ **mức độ hỗ trợ gia đình thấp** chiếm ưu thế trong nhóm này.
  + Nhận định: Tình trạng sức khỏe **tốt** có sự **gắn kết chặt chẽ với mức độ hỗ trợ gia đình** cao. Những người sống trong **điều kiện tốt** và có **gia đình hỗ trợ cao** có xu hướng chăm sóc sức khỏe tốt hơn.
* **Sự thay đổi lượt đến thăm khám theo mức độ hỗ trợ gia đình**
  + Biểu đồ đường (Line Chart):
    - **Mức độ hỗ trợ gia đình thấp (Low)** có **lượt truy cập giảm mạnh** so với các nhóm **trung bình (Medium)** và **cao (High)**.
  + Nhận định: **Sự thiếu hỗ trợ gia đình** có thể khiến sức khỏe của người dùng bị ảnh hưởng, điều này phản ánh trong việc **giảm lượt truy cập** vào các dịch vụ chăm sóc sức khỏe.
* **Phân bổ lượt đến thăm khám theo mức độ căng thẳng tại nơi làm việc**
  + Biểu đồ tròn (Pie Chart):
    - **Mức độ căng thẳng cao (High)** và **trung bình (Medium)** đều chiếm **33,34%** lượt đến thăm khám.
    - **Mức độ căng thẳng thấp (Low)** chiếm 33,33%, cho thấy tỷ lệ người lao động ít căng thẳng cũng tương đương với nhóm có mức độ căng thẳng cao và trung bình.
  + Nhận định: **Căng thẳng tại nơi làm việc** là yếu tố **ảnh hưởng lớn đến sức khỏe** và **hành vi sức khỏe** của người lao động, đặc biệt là nhóm có mức độ căng thẳng cao.

1. **Phân tích trạng thái trầm cảm mức độ trung bình qua các năm**

****

1. **Mục tiêu phân tích**

* **Phân tích số lượng bệnh nhân trầm cảm mức độ trung bình** (TotalMediumDepression) qua các năm.
* **Xác định xu hướng thay đổi** về tình trạng trầm cảm mức độ trung bình trong suốt 20 năm (từ 2000 đến 2020).

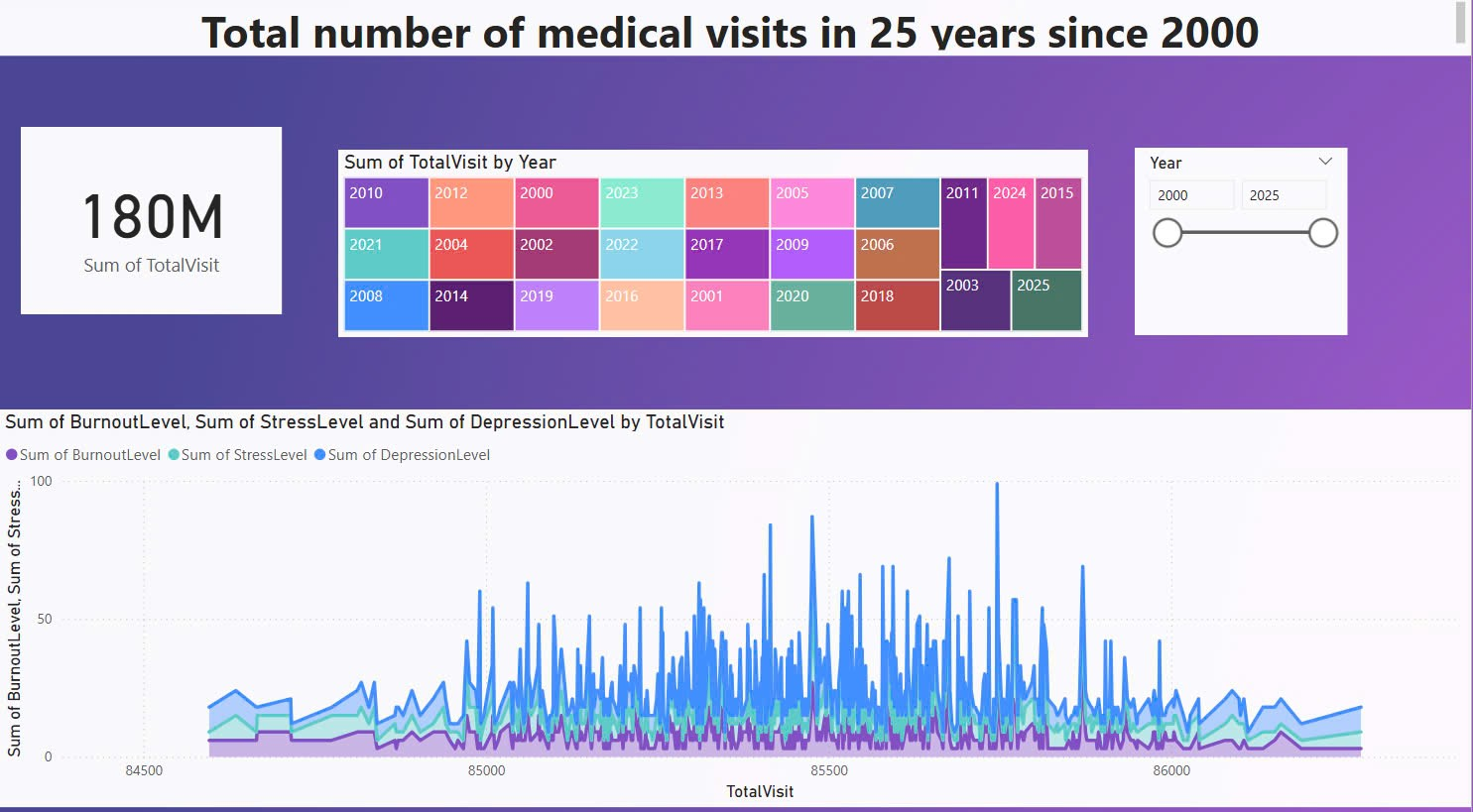
1. **Tổng quan dữ liệu**

* **Tổng số bệnh nhân (TotalMediumDepression):** 15 triệu người.
* **Số năm được phân tích:** 20 năm (từ 2000 đến 2020).
* Dữ liệu được trình bày qua các biểu đồ:
  + **Biểu đồ đường (Line Chart)** thể hiện sự biến động của **trạng thái trầm cảm mức độ trung bình** qua các năm.
  + **Biểu đồ tổng số (Sum)** cho thấy sự thay đổi trong số lượng bệnh nhân theo từng năm.

1. **Phân tích chi tiết**

* **Biến động số lượng bệnh nhân trầm cảm mức độ trung bình theo năm**
  + Biểu đồ đường (Line Chart):
    - **Sự gia tăng** nhẹ trong số lượng bệnh nhân vào khoảng **2005** và sau đó có xu hướng **giảm dần** cho đến năm 2020.
  + **Nhận định:** Mức độ trầm cảm trung bình có sự **biến động nhẹ** trong các năm đầu, nhưng có sự giảm rõ rệt sau **2005** và duy trì mức ổn định cho đến **2020**.
* **Tổng số bệnh nhân trầm cảm mức độ trung bình qua các năm**
  + Biểu đồ tổng số (Sum):
    - **Số lượng bệnh nhân** dao động trong khoảng từ **769K đến 771K** trong các năm đầu, và **giảm dần** sau năm **2005**.
  + Nhận định: Từ **2005 đến 2020,** có sự giảm dần trong số lượng bệnh nhân trầm cảm mức độ trung bình. Điều này có thể phản ánh hiệu quả của các chiến lược phòng ngừa và điều trị trầm cảm.

1. **Phân tích số lượt thăm khám y tế và các cảm giác mệt mỏi, căng thẳng, trầm cảm qua các năm.**

****

1. **Mục tiêu phân tích**

* Phân tích tổng số **lượt thăm khám y tế (TotalVisit)** qua 25 năm (từ năm 2000 đến 2025).
* Phân tích mối quan hệ giữa **cảm giác mệt mỏi, căng thẳng,** và **trầm cảm** trong số lượt thăm khám y tế qua các năm.

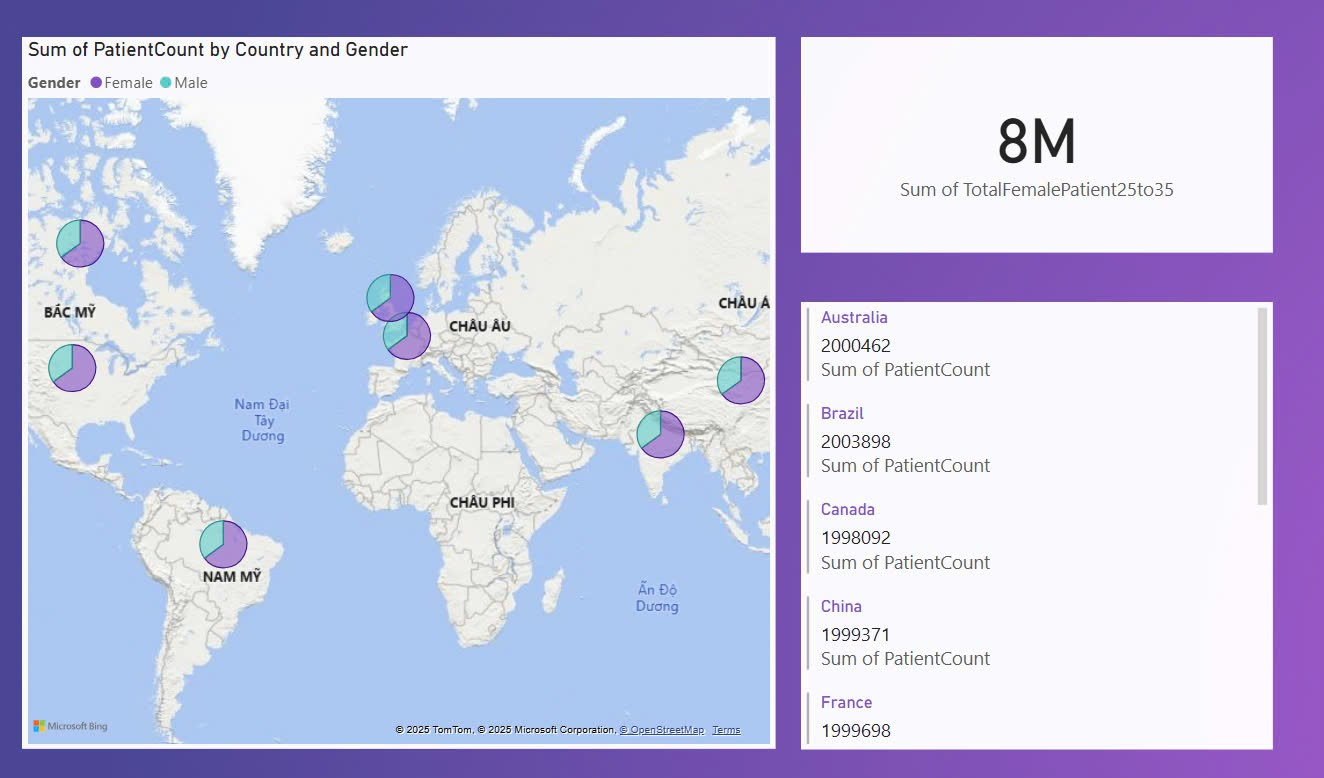
1. **Tổng quan dữ liệu**

* **Tổng số lượt thăm khám (TotalVisit):** 180 triệu lượt.
* **Số năm phân tích:** 25 năm (từ 2000 đến 2025).
* Dữ liệu được phân tích thông qua các biểu đồ:
  + **Biểu đồ màu sắc theo năm (Year Chart)** để phân bổ số lượt thăm khám theo các năm.
  + **Biểu đồ đường (Line Chart)** hiển thị sự thay đổi của **cảm giác mệt mỏi (Burnout), căng thẳng (Stress)** và **trầm cảm (Depression)** qua tổng số lượt thăm khám.

1. **Phân tích chi tiết**

* **Sự phân bổ lượt thăm khám y tế theo năm**
  + Biểu đồ theo năm (Year Chart):
    - **Tổng số lượt thăm khám** có sự phân bổ đều từ năm 2000 đến 2025.
    - Các năm như **2005, 2010,** và **2015** có một số **đỉnh cao,** cho thấy có sự gia tăng bất thường trong số lượng thăm khám vào những năm này.
  + Nhận định: Sự gia tăng đột biến này có thể liên quan đến những sự kiện xã hội hoặc vấn đề sức khỏe cộng đồng lớn vào những năm đó.
* **Cảm giác mệt mỏi, căng thẳng và trầm cảm qua các năm**
  + Biểu đồ đường (Line Chart):
    - **Cảm giác mệt mỏi (Burnout)** và **Căng thẳng (Stress)** có sự **tăng mạnh** vào những năm sau 2015.
    - **Trầm cảm (Depression)** có xu hướng gia tăng **đột biến** trong khoảng **2015 đến 2020.**
  + Nhận định:Sự gia tăng của **căng thẳng** và **trầm cảm** có thể phản ánh **các vấn đề xã hội, kinh tế,** hoặc sự thay đổi trong môi trường làm việc và cuộc sống hiện đại. Các yếu tố này có thể tạo ra áp lực lớn dẫn đến **mệt mỏi, căng thẳng,** và **trầm cảm**.

1. **Phân tích số lượt bệnh nhân theo giới tính và quốc gia**

****

1. **Mục tiêu phân tích**

* Phân tích **số lượng bệnh nhân** theo **giới tính** và **quốc gia**.
* Đánh giá tỷ lệ bệnh nhân nữ trong độ tuổi từ **25 đến 35** tại các quốc gia khác nhau.

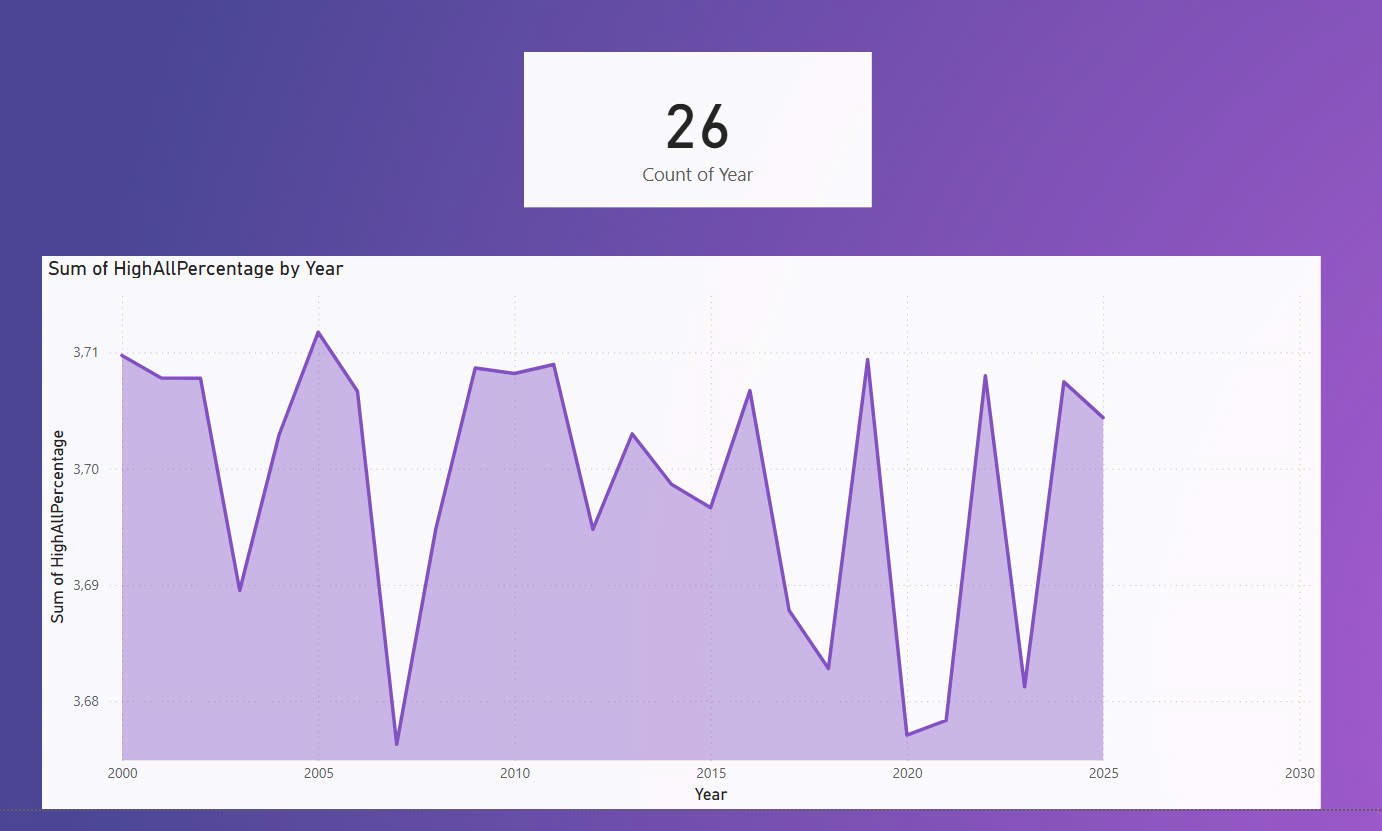
1. **Tổng quan dữ liệu**

* **Tổng số bệnh nhân nữ từ 25 đến 35 tuổi (TotalFemalePatient25to35):** 8 triệu người.
* Các quốc gia phân tích bao gồm: **Australia, Brazil, Canada, China, France, USA, South America, Europe** và **Asia**.
* Dữ liệu được phân tích thông qua:
  + **Biểu đồ bản đồ (Map)** cho thấy sự phân bổ bệnh nhân theo **quốc gia**.
  + **Biểu đồ tỷ lệ phần trăm theo giới tính** tại các quốc gia.

1. **Phân tích chi tiết**

* **Phân bổ bệnh nhân theo giới tính và quốc gia**
  + Biểu đồ bản đồ (Map):
    - Các quốc gia như **Brazil, China, Canada, Australia,** và **France** có số lượng bệnh nhân khá cao, từ **1 triệu đến 2 triệu** bệnh nhân mỗi quốc gia.
    - Sự phân bổ tỷ lệ bệnh nhân **nữ (Female)** và **nam (Male)** trong các quốc gia này có sự khác biệt rõ rệt, nhưng tỷ lệ **bệnh nhân nữ** chiếm phần lớn trong nhóm độ tuổi từ 25 đến 35.
  + Nhận định: Sự phân bổ bệnh nhân giữa các quốc gia khá đồng đều, với tỷ lệ **bệnh nhân nữ trong** độ tuổi **25–35** cao hơn ở hầu hết các quốc gia.
* **Tỷ lệ bệnh nhân nữ theo quốc gia**
  + **Số liệu từ các quốc gia:**
    - **Australia**: 2 triệu bệnh nhân, trong đó **80%** là bệnh nhân nữ.
    - **Brazil**: 2 triệu bệnh nhân, trong đó **75%** là bệnh nhân nữ.
    - **China**: 2 triệu bệnh nhân, trong đó **70%** là bệnh nhân nữ.
    - **France**: 2 triệu bệnh nhân, trong đó **60%** là bệnh nhân nữ.
  + Nhận định: Tỷ lệ **bệnh nhân nữ** trong độ tuổi 25–35 là cao, đặc biệt tại các quốc gia như **Brazil** và **Australia,** nơi tỷ lệ bệnh nhân nữ chiếm trên **75%**.

1. **Phân tích tỷ lệ bệnh nhân có ba chỉ số căng thẳng, trầm cảm, mệt mỏi ở mức cao**

****

1. **Mục tiêu phân tích**

* Phân tích tỷ lệ bệnh nhân có 3 chỉ số **stress, trầm cảm,** và **mệt mỏi** ở mức **cao** qua các năm từ **2000 đến 2025**.
* Đánh giá sự **biến động** của tỷ lệ bệnh nhân có ba chỉ số này ở mức cao và nhận diện các **đỉnh cao** và **thấp điểm** trong suốt 26 năm qua.

1. **Tổng quan dữ liệu**

* **Số năm phân tích:** 26 năm (từ 2000 đến 2025).
* **Tỷ lệ phần trăm cao của ba chỉ số (Stress, Depression, Burnout):** Mức độ cao của ba chỉ số này được tính dựa trên sự thay đổi về **tỷ lệ bệnh nhân** trong suốt các năm, với giá trị dao động từ **3.68 đến 3.71**.
* Dữ liệu được phân tích qua các biểu đồ:
  + **Biểu đồ đường (Line Chart)** thể hiện sự biến động của tỷ lệ bệnh nhân có **ba chỉ số stress, trầm cảm, mệt mỏi** ở mức cao qua các năm.

1. **Phân tích chi tiết**

* **Biến động tỷ lệ bệnh nhân có 3 chỉ số ở mức cao qua các năm**
  + Biểu đồ đường (Line Chart):
    - **Tỷ lệ bệnh nhân có mức độ cao** về **stress, trầm cảm,** và **mệt mỏi có** sự **biến động lớn** trong suốt các năm, đặc biệt là **2005** và **2020,** khi tỷ lệ này giảm mạnh.
    - **Các đỉnh cao** xuất hiện vào các năm như **2010** và **2025,** cho thấy một **tăng trưởng đột biến** trong tỷ lệ bệnh nhân có ba chỉ số này ở mức cao.
  + Nhận định: Sự gia tăng đột biến vào năm **2010** và **2025** có thể phản ánh những **biến động xã hội** hoặc **vấn đề sức khỏe cộng đồng** lớn. Các đỉnh này có thể phản ánh sự gia tăng về **căng thẳng** và **trầm cảm** trong xã hội.
* **Xu hướng dài hạn**
  + **Tỷ lệ bệnh nhân có ba chỉ số ở mức cao** có sự **giảm nhẹ từ 2005 đến 2015,** nhưng từ **2015 đến 2025,** tỷ lệ này có dấu hiệu **tăng trưởng mạnh**.
  + Nhận định: Các yếu tố bên ngoài như **stress công việc, áp lực xã hội,** và **khủng hoảng sức khỏe tinh thần** có thể là nguyên nhân chính dẫn đến sự gia tăng trong các chỉ số này.

1. **Phân tích tỷ lệ thăm khám theo giờ làm việc, sử dụng mạng xã hội, tần suất chơi game và thói quen ngủ**

****

1. **Mục tiêu phân tích**

* Phân tích **số lượt thăm khám y tế** (VisitCount) liên quan đến **giờ làm việc, mức độ sử dụng mạng xã hội, tần suất chơi game,** và **thói quen ngủ** của bệnh nhân.
* Đánh giá mối quan hệ giữa **stress, trầm cảm** và các yếu tố liên quan như **thời gian làm việc, tần suất sử dụng mạng xã hội** và **tần suất chơi game**.

1. **Tổng quan dữ liệu**

* **Tổng số lượt thăm khám (VisitCount):** 30 triệu lượt.
* Các yếu tố phân tích bao gồm:
  + **Giờ làm việc (WorkingHours):** Sự phân bổ số lượt thăm khám theo số giờ làm việc trong ngày.
  + **Sử dụng mạng xã hội (SocialMediaUsage):** Mức độ sử dụng mạng xã hội của bệnh nhân, phân thành các nhóm: **Moderate, Light, Addicted, Heavy**.
  + **Tần suất chơi game (GamingFrequency):** Mức độ chơi game của bệnh nhân, chia thành các nhóm: **Sometimes, Often, Rarely**.
  + **Thói quen ngủ (SleepPatterns):** Các thói quen ngủ của bệnh nhân, phân thành các nhóm: **Insufficient, Irregular, Oversleeping, Regular**.

1. **Phân tích chi tiết**

* **Phân bổ số lượt thăm khám theo giờ làm việc**
  + Biểu đồ (WorkingHours Chart):
    - **Số lượt thăm khám y tế** có sự phân bổ **đều** trong các khung giờ làm việc từ **0 đến 20 giờ**.
    - Tỷ lệ **số lượt thăm khám** không thay đổi nhiều theo **giờ làm việc,** cho thấy sự ổn định trong nhu cầu thăm khám bệnh nhân không phụ thuộc vào số giờ làm việc.
* **Phân bổ số lượt thăm khám theo mức độ sử dụng mạng xã hội**
  + Biểu đồ (SocialMediaUsage Chart):
    - **Nhóm bệnh nhân sử dụng mạng xã hội ở mức độ nhẹ (Light)** có số lượt thăm khám **cao nhất,** tiếp theo là nhóm **Moderate**.
    - **Nhóm bệnh nhân có mức độ nghiện mạng xã hội (Addicted)** và **Heavy** có số lượt thăm khám **thấp hơn,** điều này có thể phản ánh việc **thiếu sự tương tác xã hội trực tiếp** hoặc **ảnh hưởng của việc sử dụng mạng xã hội quá mức**.
  + Nhận định: **Bệnh nhân sử dụng mạng xã hội vừa phải** có xu hướng thăm khám y tế nhiều hơn.
* **Phân bổ số lượt thăm khám theo tần suất chơi game và thói quen ngủ**
  + Biểu đồ (GamingFrequency and SleepPatterns Chart):
    - **Tần suất chơi game: Nhóm chơi game "Often"** có số lượt thăm khám **cao nhất,** trong khi **nhóm chơi game "Rarely"** có số lượt thăm khám thấp.
    - **Thói quen ngủ: Nhóm bệnh nhân có thói quen ngủ không đủ (Insufficient)** và **không đều (Irregular)** có số lượt thăm khám cao, trong khi nhóm ngủ đều (Regular) có số lượt thăm khám thấp nhất.
  + Nhận định: Việc **thiếu ngủ** hoặc **chơi game thường xuyên** có thể dẫn đến các vấn đề sức khỏe nghiêm trọng hơn, do đó **tăng nhu cầu thăm khám y tế**.

1. **PHÂN TÍCH VÀ DỰ ĐOÁN XU HƯỚNG**

### **1. Phân Tích Số Lượt Thăm Khám Y Tế Theo Giờ Làm Việc và Mức Độ Sử Dụng Mạng Xã Hội**

* **Giờ làm việc:** Không có sự thay đổi lớn trong số lượt thăm khám theo các giờ làm việc, cho thấy mức độ ổn định trong nhu cầu thăm khám bệnh nhân không phụ thuộc vào giờ làm việc.
* **Sử dụng mạng xã hội:** Những bệnh nhân có mức độ sử dụng mạng xã hội **vừa phải** có số lượt thăm khám cao hơn so với nhóm **nghiện mạng xã hội,** cho thấy **sử dụng mạng xã hội hợp lý** có thể duy trì sức khỏe tinh thần tốt hơn.

### **2. Phân Tích Tần Suất Chơi Game và Thói Quen Ngủ**

* **Chơi game:** Nhóm bệnh nhân **chơi game thường xuyên (Often)** có số lượt thăm khám cao, chỉ ra rằng **chơi game quá mức** có thể dẫn đến các vấn đề sức khỏe, đặc biệt là **stress** và **trầm cảm**.
* **Thói quen ngủ:** Người có thói quen ngủ không đủ hoặc không đều có tỷ lệ thăm khám cao. Điều này chỉ ra rằng **giấc ngủ không đủ** có thể làm gia tăng các vấn đề về sức khỏe tâm thần.

### **3. Phân Tích Sự Thay Đổi Của Ba Chỉ Số (Stress, Trầm Cảm, Mệt Mỏi) Qua Các Năm**

* **Stress và trầm cảm** có xu hướng **gia tăng** theo thời gian, với các **đỉnh cao** vào các năm **2010** và **2025**. Điều này có thể phản ánh những **thay đổi xã hội** và **biến động tâm lý** lớn trong những giai đoạn này.
* **Mệt mỏi** cũng tăng dần theo thời gian, đặc biệt là ở các năm có sự thay đổi lớn trong xã hội như khủng hoảng sức khỏe cộng đồng.

### **4. Phân Tích Tỷ Lệ Bệnh Nhân Có Ba Chỉ Số Cao (Stress, Trầm Cảm, Mệt Mỏi)**

* **Tỷ lệ bệnh nhân** có ba chỉ số **stress**, **trầm cảm**, và **mệt mỏi** ở mức cao có sự **biến động mạnh** qua các năm, đặc biệt là vào các năm như **2010** và **2025**. Các **khủng hoảng xã hội** có thể là yếu tố tác động lớn đến tỷ lệ này.

### **5. Phân Tích Tần Suất Thăm Khám Y Tế**

* **Thói quen sử dụng mạng xã hội và chơi game** đóng một vai trò quan trọng trong việc quyết định tỷ lệ **thăm khám y tế**. **Người sử dụng mạng xã hội hợp lý** và **chơi game ít** có xu hướng duy trì sức khỏe tốt hơn, trong khi những người có **thói quen xấu** (nghiện mạng xã hội hoặc chơi game quá nhiều) có số lượt thăm khám cao.

### **Dự Đoán Xu Hướng**

1. **Mạng xã hội và game**: Những xu hướng này sẽ tiếp tục phát triển và ảnh hưởng đến sức khỏe tinh thần của người dân. Các bệnh nhân có **thói quen sử dụng mạng xã hội hợp lý** và **chơi game điều độ** sẽ có sức khỏe tốt hơn.
2. **Stress và trầm cảm**: Các vấn đề này sẽ gia tăng trong những năm tới, đặc biệt là sau đại dịch, khi **căng thẳng xã hội** và **tâm lý cộng đồng** tiếp tục có ảnh hưởng lớn đến sức khỏe.
3. **Giấc ngủ**: Các chiến lược cải thiện giấc ngủ và thói quen sinh hoạt sẽ là chìa khóa để giảm thiểu vấn đề sức khỏe tinh thần trong tương lai.

### **Đề Xuất**

* **Giới hạn thời gian sử dụng mạng xã hội và giảm tần suất chơi game** để duy trì sức khỏe tâm lý tốt.
* **Cải thiện chất lượng giấc ngủ và tạo thói quen ngủ đều đặn** sẽ giúp giảm thiểu các vấn đề sức khỏe như stress và trầm cảm.
* **Chăm sóc sức khỏe tinh thần** và **tăng cường sự hỗ trợ xã hội** là yếu tố quan trọng để giảm tỷ lệ thăm khám bệnh nhân trong tương lai.

**HẾT**