

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA CÔNG NGHỆ THÔNG TIN**



**Logo

Description automatically generated**

**MÔN TRỰC QUAN HÓA DỮ LIỆU**

**BÁO CÁO LAB 2:**

**SỬ DỤNG TABLEAU ĐỂ TRỰC QUAN HÓA DỮ LIỆU**

***Thứ bảy, ngày 28 tháng 5 năm 2022***

**

1. **Thông tin nhóm:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Họ và tên** | **MSSV** | **Công việc** | **Mức độ hoàn thành** |
| Hồ Hữu Ngọc | 19120602 | * Viết báo cáo * Tìm hiểu về tableau và vẽ biểu đồ trực quan sử dụng tableau | **100%** |
| Nguyễn Phát Minh | 19120586 |  | **100%** |
| Võ Trọng Phú | 19120328 | * Tìm hiểu về Tableau, * Chạy mô hình máy học * Viết báo cáo các phần trên | **100%** |

1. **Mức độ hoàn thành tổng thể mỗi yêu cầu:**

|  |  |
| --- | --- |
| **Công việc** | **Mức độ hoàn thành** |
| Tìm hiểu về Tableau | **100%** |
| Áp dụng Tableau để trực quan dữ liệu (một ngày và nhiều ngày) | **100%** |
| Sử dụng các kỹ thuật được giới thiệu trong bài Manipulate View, Facet, Reduce, Embed để trình diễn trên Tableau với dữ liệu Woldometer | **100%** |
| Chạy các thuật toán máy học đơn giản để hiểu them về dữ liệu | **100%** |

1. **Tìm hiểu về Tableau:**
2. ***Giới thiệu về Tableau:***

* Tableau là một phần mềm được sử dụng để hỗ trợ phân tích và trực quan hóa dữ liệu, giúp người dùng tạo ra các biểu đồ trực quan dựa trên nguồn dữ liệu.
* Được sử dụng nhiều trong ngành Business Intelligence (BI).
* Tableau có nhiều ưu điểm so với các ứng dụng khác, người sử dụng có thể dễ dàng chuyển đổi những dữ liệu dạng thô sang những hình ảnh trực quan bắt mắt và dễ hình dung, ngoài ra nó có thể hoạt đông với bất kì loại dữ liệu nào (có cấu trúc, phi cấu trúc, bán cấu trúc).
* Người dùng Tableau không cần nền tảng kĩ thuật để sử dụng công cụ này.

1. ***Các tính năng hỗ trợ của Tableau và ví dụ:***
2. ***Khả năng kết nối với dữ liệu:***
   * Tableau có thể kết nối với nhiều loại dữ liệu, môi trường dữ liệu khác nhau như: File (Excel, Text file, JSON file, PDF file,…), Server (SQL Server, Oracle, MySQL,..) hoặc một Data Sources đã được lưu trước đó.
   * Ví dụ: ta thử kết nối Tableau đến một file Excel đã lưu trong máy.

A screenshot of a computer

Description automatically generated

1. ***Tính năng kéo thả:***

* Giao diện điều khiển của Tableau giúp người dùng có thể dễ dàng thực hiện thao tác kéo thả các trường, các mục điều khiển theo ý mình.
* Ví dụ: kéo thả trường Order vào vùng hoạt động.

Graphical user interface

Description automatically generated

1. ***Tính năng mô phỏng dữ liệu (Data Visualization):***

* Khi dữ liệu đã được kết nối, bằng cách kéo thả mục cần trực quan ta có thể tùy chọn trực quan hóa dữ liệu dưới nhiều tùy chọn biểu đồ khác nhau (góc bên phải màn hình)
* Ví dụ: biểu đồ bar chart trực quan Sales theo từng quốc gia.

A screenshot of a computer

Description automatically generated

1. ***Tính năng xây dựng bản đồ:***

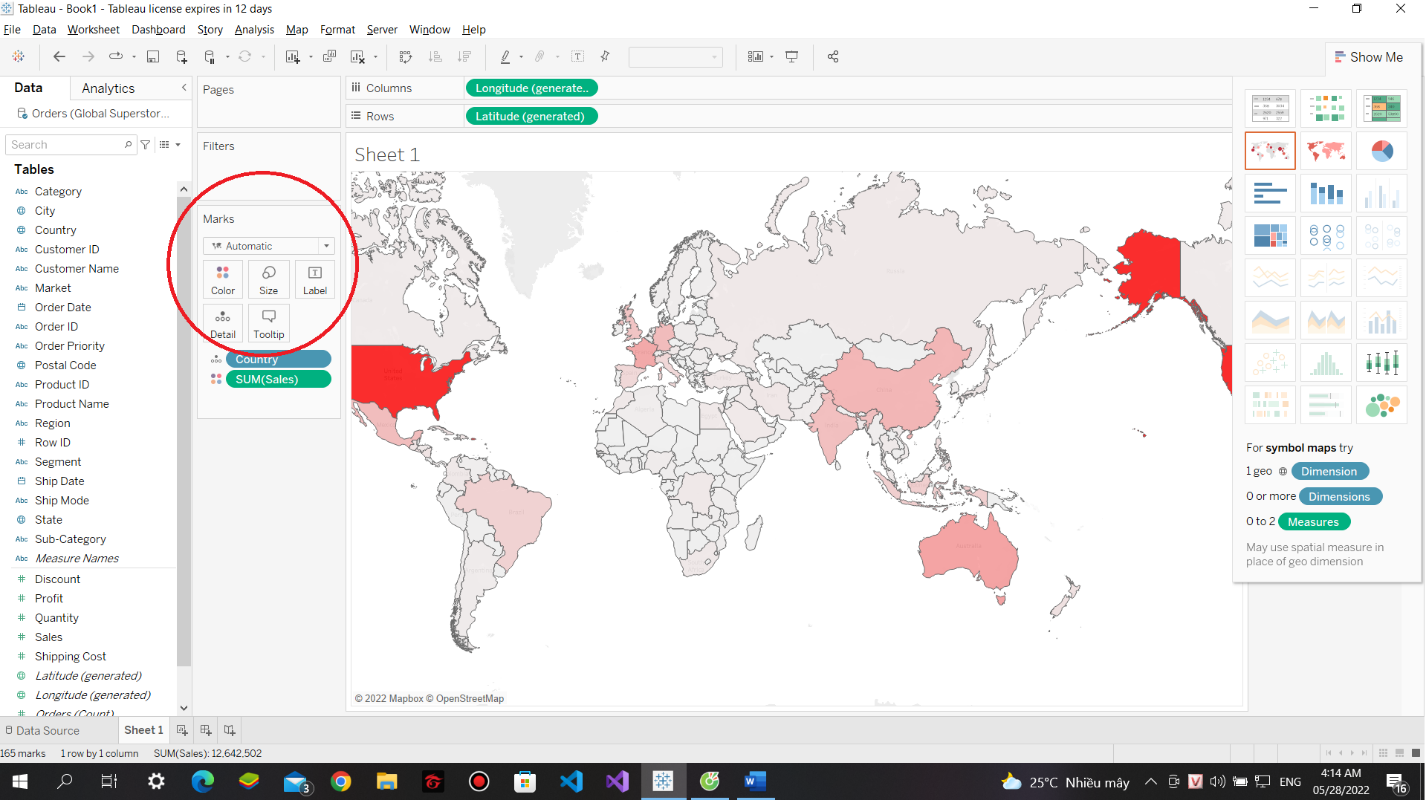
* Ta có thể xây dựng một hình thức trực quan mới, đó là vẽ bản đồ địa lí cho dữ liệu, vẫn là những lựa chọn ở góc bên phải màn hình.
* Ví dụ: trực quan Sales theo từng quốc gia bằng bản đồ.

Map

Description automatically generated

1. ***Tính năng tùy chỉnh các biểu đồ:***

* Ta có thể tùy chỉnh kích thước, màu sắc, label,...của biểu đồ ở bảng điều khiển trong hình.
* Ví dụ: thay đổi màu sắc của bản đồ trực quan ở trên sang màu đỏ.



1. ***Các tính năng khác:***

* Các tính năng hỗ trợ khác như chia sẻ, in ra file PDF, export sang powerpoint,… cũng được hỗ trợ trong Tableu ở mục tùy chọn File.
* Ví dụ: chọn File -> Share để dùng tính năng chia sẻ.

Graphical user interface, application, map

Description automatically generated

1. **Áp dụng Tableau để trực quan dữ liệu (một ngày và nhiều ngày):**
2. **Top 10 nước có số lượng ca nhiễm mới trong ngày.**

Chart

Description automatically generated

* Data được chọn là ngày 21/04/2022 sử dụng hai trường là(Country và New Case) lấy ra top 50 nước có số lượng ca nhiễm mới mỗi ngày và vẽ bằng barchar. Nhìn vào biểu đồ có thể thấy số lượng ca nhiễm vẫn tăng hàng ngày. Nhưng số ca mỗi ngày tương đối ít. Điển hình là nước Bhutan cũng không quá 1000 ca/ngày

1. **Tỉ lệ số ca chết trong ngày của top 5 nước có số người chết nhiều nhất trong ngày.**

* Chart, pie chart

  Description automatically generated
* Dữ liệu của ngày 21/04/2022 sử dụng hai trường là (Country và New Death) cho thấy 5 nước có số lượng người chết trong ngày nhiều nhất do Covid 19. Trong ngày này thì Finland có số người chết nhiều nhất, tỷ lệ hơn hẳn các nước còn lại.

1. **Chạy một số mô hình học máy đơn giản để hiểu thêm về dữ liệu:**
2. ***Thuật toán giảm chiều Principal Component Analysis (PCA):***

* Thuật toán PCA dùng để chuyển n chiều sang dữ liệu m chiều (m < n) mà vẫn giữ được nhiều thông tin nhất có thể.
* Các bước thực hiện như sau:
* Load dữ liệu, lấy các dữ liệu số và bỏ đi cột ‘Country, Other’
* Chuẩn hóa dữ liệu (dùng hàm fit\_transform)
* Dùng PCA giảm chiều dữ liệu xuống 2 components, lúc này ta sẽ được kết quả giảm chiều từ 13 cột còn 2 cột:

Text

Description automatically generated

* Thêm lại cột ‘Country, Other’ vào data mới.

Graphical user interface, text, application, email

Description automatically generated

* Tiến hành trực quan dữ liệu cuối cùng bằng Scatter plot, để thấy rõ được tương quan giữa các nước, ta sẽ bỏ đi dòng World:

Chart, scatter chart

Description automatically generated

Chỉ có một chấm tròn duy nhất ở phía góc trên phải có giá trị xấp xỉ (5,8), đa số chấm tròn các nước đều tập trung ở góc trái và có Principal Components 1 âm.

1. ***Thuật toán bình phương nhỏ nhất Ordinary Least Squares (OLS):***