ĐẠI HỌC UEH

TRƯỜNG CÔNG NGHỆ VÀ THIẾT KẾ

KHOA CÔNG NGHỆ THÔNG TIN KINH DOANH

ĐỒ ÁN CUỐI KỲ

BIỂU DIỄN TRỰC QUAN DỮ LIỆU

Đề tài: Biểu diễn trực quan dữ liệu về Chỉ số Hạnh phúc của các quốc gia trên thế

Thành viên:

Đặng Đại Lợi – 31201021601

Nguyễn King - 31211023531

Doãn Phương Hà My - 31211027649

Võ Ngọc Mỹ Kim - 31211027646

Giảng viên hướng dẫn: TS. Nguyễn An Tế

Thành phố Hồ Chí Minh, ngày 26 tháng 11 năm 2023

MỤC LỤC

MỤC LỤC BIỂU ĐỒ

LỜI CẢM ƠN

Mỗi người trong thời đại hiện nay luôn tồn tại những áp lực riêng, áp lực từ công việc, áp lực từ học tập, áp lực về các định kiến của xã hội,… Trong những năm qua, nhằm để nâng cao chất lượng cuộc sống, bên cạnh mối quan tâm về sức khỏe thể chất và sức khỏe tài chính, mọi người ngày càng có nhận thức và đề cao sức khỏe tinh thần. Sức khỏe tinh thần ngày nay càng được quan tâm, chỉ số hạnh phúc của từng quốc gia ngày càng được chú ý. Chỉ số hạnh phúc được nhiều người xem là một cách đo lường chất lượng cuộc sống, nó không chỉ bao gồm các chỉ số liên quan đến các yếu tố về vật chất như thu nhập hay GDP của một quốc gia mà còn bao gồm các yếu tố khác liên quan đến tâm lý, xã hội và sức khỏe.

Nhóm chọn bộ dữ liệu ‘World Happiness 2017’ để có thể phân tích, khai thác và truyền tải thông tin từ các yếu tố liên quan đến chỉ số hạnh phúc một cách trực quan nhất. Mục đích của việc phân tích bộ dữ liệu này là diễn giải các yếu tố làm ảnh hưởng đến range điểm chỉ số hạnh phúc, và chỉ số hạnh phúc của từng châu lục theo 4 nhóm điểm từ ‘Cao’, ‘Trung bình - Cao’, ‘Trung bình - Thấp’, ‘Thấp’, để truyền tải đến cho những người sử dụng dữ liệu có thể nắm rõ hơn về các yếu tố.

Quá trình thực hiện đồ án ngoài việc giúp chúng em có thể vận dụng được các kiến thức được học để phân tích và trực quan hóa dữ liệu mà còn giúp chúng em có thêm sự hiểu biết về đề tài chúng em đang thực hiện - yếu tố ảnh hưởng đến chỉ số hạnh phúc.

Quá trình làm đồ án môn học, chúng em vẫn còn nhiều hạn chế về kiến thức, sai sót và chưa tối ưu về mặt kỹ thuật. Nhóm mong sẽ nhận được phản hồi của thầy để có thể cải thiện hơn về mặt kiến thức chuyên môn.

Nhóm chúng em xin được gửi lời cảm ơn về sự tận tâm, nhiệt huyết của thầy thông qua các bài giảng trên lớp, thầy đã truyền tải các kiến thức bổ ích, quan trọng và cung cấp các tài liệu để giúp chúng em có đủ kiến thức và kỹ năng cần thiết để có thể hoàn thành đồ án của môn học “Biểu diễn dữ liệu trực quan”.

Nhóm chúng em xin chân thành cảm ơn thầy.

CHƯƠNG I: TỔNG QUAN ĐỀ TÀI

1.1. Giới thiệu đề tài

Mức độ Hạnh phúc là một trong những yếu tố khẳng định sự thành công của không chỉ một cá nhân mà còn của cả một quốc gia. Chỉ số Hạnh phúc của một quốc gia có thể liên quan tới nhiều yếu tố xã hội và kinh tế. Mức độ Hạnh phúc phản ánh sự hiệu quả của các chính sách hiện có của một chính quyền.

Việc phân tích Chỉ số Hạnh phúc và các yếu tố liên quan không chỉ giúp ích cho một cá nhân để tham khảo mà còn giúp ích cho các lãnh đạo quốc gia và toàn thế giới.

1.2. Mục tiêu nghiên cứu

Phân tích bộ dữ liệu về Chỉ số Hạnh phúc Toàn cầu để mô tả tình hình sức khỏe tinh thần của các quốc gia theo lục địa. Ngoài ra, phân tích các yếu tố có liên quan đến Chỉ số Hạnh phúc, từ đó đưa ra lời khuyên về chính sách phù hợp. Mô tả quá trình rút trích thông tin và trực quan hóa để rút ra các kết luận để hiểu rõ hơn về bộ dữ liệu sử dụng.

1.3. Phương pháp nghiên cứu:

Thống kê: sử dụng các công thức và kỹ thuật trong Thống kê để tính toán các chỉ số phù hợp cho việc trực quan hóa và đánh giá.

Học máy: sử dụng các mô hình Học máy để giải quyết bài toán]Phân cụm, Hồi quy

Trực quan biểu đồ: lựa chọn các biểu đồ phù hợp với loại dữ liệu cần phân tích, loại thông tin cần mô tả giúp người đọc hình dung và đánh giá dễ dàng hơn.

1.4. Tài nguyên sử dụng

Ngôn ngữ lập trình: Python

Các thư viện sử dụng: Pandas, Numpy, Seaborn, Matplotlib, Plotly,…

Bộ dữ liệu được lấy từ báo cáo “Hạnh phúc Toàn cầu”, được xuất bản bởi Mạng lưới Giải pháp Phát triển Bền vững thuộc Liên Hợp Quốc. Báo cáo này chủ yếu lấy dữ liệu từ Khảo sát Toàn cầu Gallup. Khảo sát này dùng 100 câu hỏi Toàn cầu về các lĩnh vực như luật pháp, thức ăn và nơi ở, cơ sở vật chất, … với đối tượng là 1000 cư dân mỗi quốc gia.

CHƯƠNG II: BỘ DỮ LIỆU & TIỀN XỬ LÝ

2.1. Nhập dữ liệu

2.1.1. Mô tả bộ dữ liệu

Bộ dữ liệu "World Happiness Report - 2017" được thu thập bởi Gallup World Poll và được xuất bản bởi The United Nations Sustainable Development Solutions Network.

Bộ dữ liệu này là một cuộc khảo sát về tình trạng hạnh phúc toàn cầu và xếp hạng 155 quốc giá theo mức độ hạnh phúc trong cuộc sống và làm việc của các công dân của mỗi nước.

2.1.2. Ý nghĩa thuộc tính

Để hiểu rõ hơn về bộ dữ liệu, trước tiên ta cần phải nắm rõ tất cả các thuộc tính có trong bộ dữ liệu nguyên bản.

Tên thuộc tính	**Mô tả**	**Ghi chú**
Country	Tên quốc gia	Bao gồm 155 quốc gia và vùng lãnh thổ.
Happiness.Rank	Xếp hạng hạnh phúc của quốc gia	Quốc gia hạnh phúc nhất được xếp hạng 1
Happiness.Score	Điểm số hạnh phúc của quốc gia	Được tính toán dựa trên các câu hỏi đánh giá cuộc sống trong cuộc khảo sát. Thang điểm từ 0 đến 10, trong đó điểm 10 là quốc gia hạnh phúc nhất.
Whisker.high/ Whisker.low	Các giá trị này tạo thành khoảng tin cậy 95% cho điểm số hạnh phúc	

Economy..GDP.per .Capita.	Sản lượng quốc nội trên đầu người	Đại diện cho mức độ sản xuất kinh tế của một quốc gia
Family	Đánh giá về hỗ trợ xã hội	
Health..Life.Expectancy	Tuổi thọ trung bình	
Freedom	Mức độ tự do cá nhân của mỗi quốc gia	
Generosity	Mức độ rộng lượng	Thường được đo lường bằng tỷ lệ quyên góp
Trust	Mức độ tin tưởng và sự nhận thức về tham nhũng trong chính phủ	
Dystopia.Residual	Là điểm số mà mỗi quốc gia phải vượt qua để có điểm số hạnh phúc (cao hơn Dystopia), nó được sử dụng làm điểm chuẩn hồi quy.	Dystopia là một quốc gia giả tưởng có các giá trị thấp nhất trên thế giới cho mỗi trong sáu yếu tố - sản xuất kinh tế, hỗ trợ xã hội, tuổi thọ, tự do, không có tham nhũng và sự rộng lượng

2.2. Điều chỉnh định dạng dữ liệu

Xét thấy tên các thuộc tính khá dài dòng sẽ gây khó khăn hay dễ nhầm lẫn trong quá trình phân tích. Vì thế ta nên thay đổi ngắn gọn tên một số thuộc tính như sau:

Trước khi đổi	**Sau khi đổi**
Country	Country
Happiness.Rank	Rank
Happiness.Score	Score
Whisker.high/ Whisker.low	Whisker.high/ Whisker.low
Economy..GDP.per.Capita.	GDP
Family	Family
Health..Life.Expectancy	Life Expectancy
Freedom	Freedom
Generosity	Generosity
Trust	Trust
Dystopia.Residual	Dystopia

2.3. Xử lý giá trị bị thiếu

Trong quá trình phân tích chi tiết, ta cần xử lý các vấn đề có thể ảnh hưởng đến chất lượng bộ dữ liệu, dẫn đến việc đưa ra kết quả thiếu chính xác. Một trong số các vấn đề đó là các giá trị bị thiếu.

Để xử lý các giá trị bị thiếu trước tiên, ta kiểm tra thông tin của bộ dữ liệu hiện tại:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 155 entries, 0 to 154
Data columns (total 12 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   Country          155 non-null    object 
 1   Rank             155 non-null    int64  
 2   Score            155 non-null    float64
 3   Whisker.high     155 non-null    float64
 4   Whisker.low      155 non-null    float64
 5   GDP              155 non-null    float64
 6   Family           155 non-null    float64
 7   Life Expectancy  155 non-null    float64
 8   Freedom          155 non-null    float64
 9   Generosity       155 non-null    float64
 10  Trust            155 non-null    float64
 11  Dystopia         155 non-null    float64
dtypes: float64(10), int64(1), object(1)
memory usage: 14.7+ KB
```

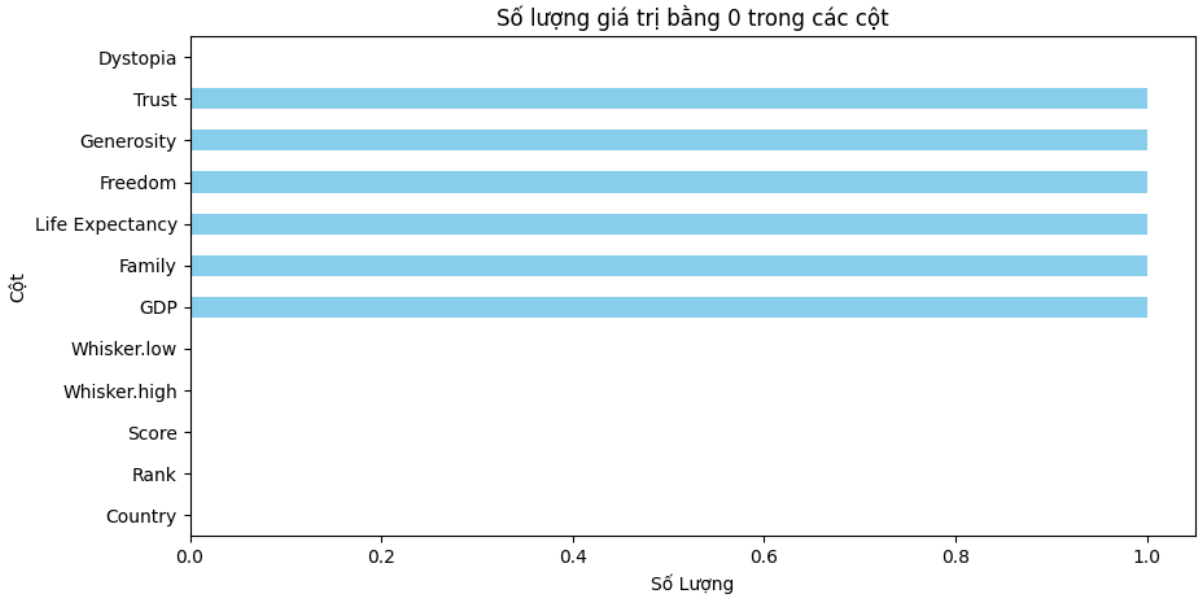
Nhận xét: Theo như kết quả cho thấy rằng bộ dữ liệu không hề có giá trị NaN.

Tuy nhiên, sau khi nhìn tổng thể cả bộ dữ liệu nhận rằng, các giá trị NaN được thay thế bằng 0.

```
1 df.tail(20)
```

	Country	Rank	Score	Whisker.high	Whisker.low	GDP	Family	Life Expectancy	Freedom	Generosity	Trust	Dystopia
135	Malawi	136	3.970	4.077479	3.862521	0.233442	0.512569	0.315090	0.466915	0.287170	0.072712	2.081786
136	Chad	137	3.936	4.034712	3.837289	0.438013	0.953856	0.041135	0.162342	0.216114	0.053582	2.071238
137	Zimbabwe	138	3.875	3.978700	3.771300	0.375847	1.083096	0.196764	0.336384	0.189143	0.095375	1.597970
138	Lesotho	139	3.808	4.044344	3.571656	0.521021	1.190095	0.000000	0.390661	0.157497	0.119095	1.429835
139	Angola	140	3.795	3.951642	3.638358	0.858428	1.104412	0.049869	0.000000	0.097926	0.069720	1.614482
140	Afghanistan	141	3.794	3.873661	3.714338	0.401477	0.581543	0.180747	0.106180	0.311871	0.061158	2.150801
141	Botswana	142	3.766	3.874123	3.657877	1.122094	1.221555	0.341756	0.505196	0.099348	0.098583	0.377914
142	Benin	143	3.657	3.745784	3.568217	0.431085	0.435300	0.209930	0.425963	0.207948	0.060929	1.885631
143	Madagascar	144	3.644	3.714319	3.573681	0.305809	0.913020	0.375223	0.189197	0.208733	0.067232	1.584613
144	Haiti	145	3.603	3.734715	3.471285	0.368610	0.640450	0.277321	0.030370	0.489204	0.099872	1.697168
145	Yemen	146	3.593	3.692750	3.493250	0.591683	0.935382	0.310081	0.249464	0.104125	0.056767	1.345601
146	South Sudan	147	3.591	3.725539	3.456462	0.397249	0.601323	0.163486	0.147062	0.285671	0.116794	1.879567
147	Liberia	148	3.533	3.653756	3.412244	0.119042	0.872118	0.229918	0.332881	0.266550	0.038948	1.673286
148	Guinea	149	3.507	3.584428	3.429572	0.244550	0.791245	0.194129	0.348588	0.264815	0.110938	1.552312
149	Togo	150	3.495	3.594038	3.395962	0.305445	0.431883	0.247106	0.380426	0.196896	0.095665	1.837229
150	Rwanda	151	3.471	3.543030	3.398970	0.368746	0.945707	0.326425	0.581844	0.252756	0.455220	0.540061
151	Syria	152	3.462	3.663669	3.260331	0.777153	0.396103	0.500533	0.081539	0.493664	0.151347	1.061574
152	Tanzania	153	3.349	3.461430	3.236570	0.511136	1.041990	0.364509	0.390018	0.354256	0.066035	0.621130
153	Burundi	154	2.905	3.074690	2.735310	0.091623	0.629794	0.151611	0.059901	0.204435	0.084148	1.683024
154	Central African Republic	155	2.693	2.864884	2.521116	0.000000	0.000000	0.018773	0.270842	0.280876	0.056565	2.066005

Vì thế ta cần xem có bao nhiêu giá trị bằng 0 trong bộ dữ liệu. Ở đây ta dùng biểu đồ Histogram để dễ dàng nhìn thấy có bao nhiêu giá trị bằng 0, cùng với đó là so sánh số lượng các giá trị bằng 0 trong các cột với nhau.



Biểu đồ 1.1 Histogram graph - Số lượng các giá trị bằng 0 trong các cột

Nhận xét: Trong bộ dữ liệu này có 6 cột có giá trị bằng 0 là Trust, Generosity, Freedom, Life Expectancy, Family và GDP, trong đó số lượng ở mỗi cột là bằng nhau và chỉ có một giá trị.

Để khắc phục điều này có một số phương pháp như xoá bỏ hàng chứa giá trị bằng 0, thay thế chúng bằng giá trị mean hoặc median, hoặc dùng những mô hình dự đoán.

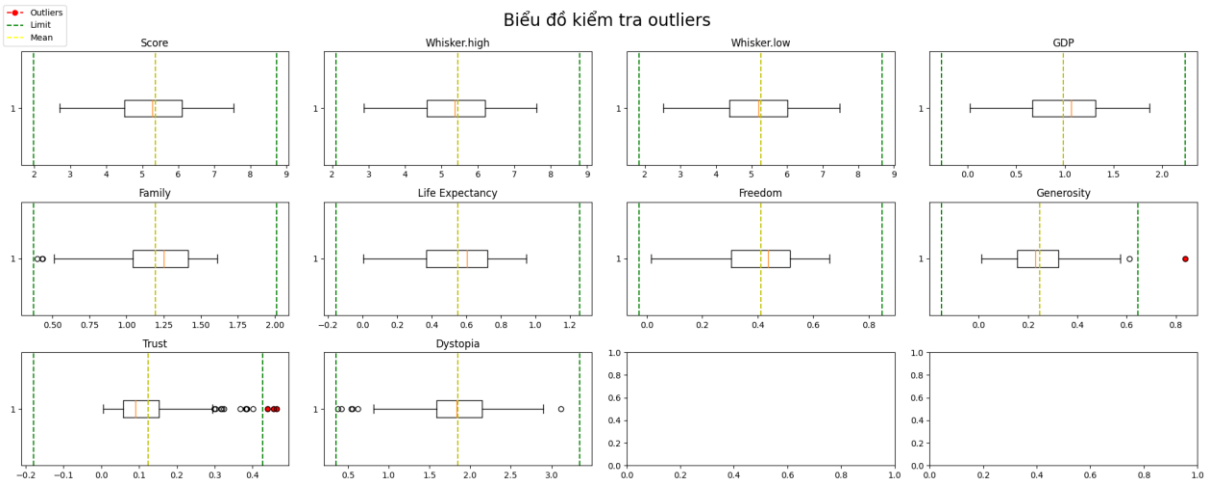
Xét trong trường hợp của bộ dữ liệu "World Happiness Report - 2017" này, ta có thể thấy đây là một bộ dữ liệu xếp hạng từ cao xuống thấp theo thang điểm hạnh phúc. Việc xóa bỏ hàng chứa giá trị bằng 0 sẽ làm mất dữ liệu của quốc gia đó nên phương án này đương nhiên không thích hợp. Bên cạnh đó, phương án điền mean và median có thể không phản ánh chính xác thứ tự của dữ liệu. Cuối cùng, nhóm đưa ra phương án ít sai lệch nhất là dùng interpolation - sử dụng các giá trị gần đó để thay thế cho giá trị bị thiếu.

2.4. Xử lý giá trị ngoại lai

Bên cạnh giá trị thiếu, nhân tố tiếp theo gây ảnh hưởng đến chất lượng dữ liệu đó là giá trị ngoại lai (outlier).

Ta sử dụng quy tắc 3-sigma để xác định các giá trị ngoại lai. Ở đây nhóm chọn biểu diễn tổng thể các giá trị ngoại lai bằng Boxplot và in ra chi tiết dòng và cột chứa giá trị ngoại lai.

Boxplot thể hiện phân phối của dữ liệu, biểu đồ này cho ta thấy các đặc điểm như độ dàn trải của dữ liệu, sự đối xứng và các giá trị ngoại lai của bộ dữ liệu.



Biểu đồ 2.2. Box Plot graph - Biểu đồ kiểm tra outliers

Nhận xét: Biểu đồ cho thấy chỉ biến Trust có 3 giá trị và Generosity có 1 giá trị ngoại lai. Bên cạnh đó, cả 4 giá trị này đều nằm bên phải của biểu đồ, tức là các giá trị ngoại lai này có giá trị cao hơn so với phần lớn dữ liệu hiện tại.

Chi tiết về 4 giá trị ngoại lai này:

```
Dòng [113] cột "Generosity" có outlier là: 0.838075160980225
Dòng [25] cột "Trust" có outlier là: 0.46430778503418
Dòng [34] cột "Trust" có outlier là: 0.439299255609512
Dòng [150] cột "Trust" có outlier là: 0.455220013856888
```

	Country	Rank	Score	Whisker.high	Whisker.low	GDP	Family	Life Expectancy	Freedom	Generosity	Trust	Dystopia
113	Myanmar	114	4.545	4.614740	4.475260	0.367111	1.123236	0.397523	0.514492	0.838075	0.188816	1.115290
25	Singapore	26	6.572	6.636723	6.507277	1.692278	1.353814	0.949492	0.549841	0.345966	0.464308	1.216362
34	Qatar	35	6.375	6.568477	6.181523	1.870766	1.274297	0.710098	0.604131	0.330474	0.439299	1.145464
150	Rwanda	151	3.471	3.543030	3.398970	0.368746	0.945707	0.326425	0.581844	0.252756	0.455220	0.540061

Nhận xét: Nhìn chung có thể thấy rằng giá trị ngoại lai của Singapore và Qatar là hợp lý, nhưng thuộc tính Generosity của Myanmar và Trust của Rwanda lại bất thường. Bởi vì:

Singapore và Qatar là hai quốc gia phát triển và có mức sống cao, dẫn theo đó là sự tin tưởng cao với chính phủ là chuyện hiển nhiên, cho nên đây không phải là giá trị ngoại lai.

Tiếp theo là Myanmar, giá trị của thuộc tính Generosity của quốc gia này cao bất thường so với 2 nước phát triển trên trong khi GDP của Myanmar cũng thấp hơn rất nhiều. Mà thuộc tính Generosity được đo lường bằng tỷ lệ quyên góp thế nên giá trị này tỷ lệ nghịch so với GDP - đại diện cho mức độ sản xuất kinh tế của một quốc gia. Hơn nữa khi xem

tổng thể của bộ dữ liệu, những quốc gia có thứ hạng gần Myanmar, giá trị thuộc tính Generosity của họ cũng khá thấp. Vì vậy ta có thể kết luận đây là giá trị ngoại lai.

Cuối cùng là giá trị của thuộc tính Trust của Rwanda, đây là quốc gia xếp gần chót bảng xếp hạng và có điểm hạnh phúc thấp. Tuy vậy, thuộc tính Trust của Rwanda khá cao, gần bằng cả hai nước có xếp hạng cao như Singapore và Qatar. Cùng với việc xem xét các nước có xếp hạng gần kề thì giá trị của Trust ở quốc gia này quả thật khá cao, nên có thể xem đây là một giá trị ngoại lai.

Để xử lý 2 giá trị ngoại lai trên ta xem chúng như 2 giá trị bị thiếu và dùng cách tương tự là dùng interpolation - sử dụng các giá trị gần đó để thay thế cho giá trị bị thiếu.

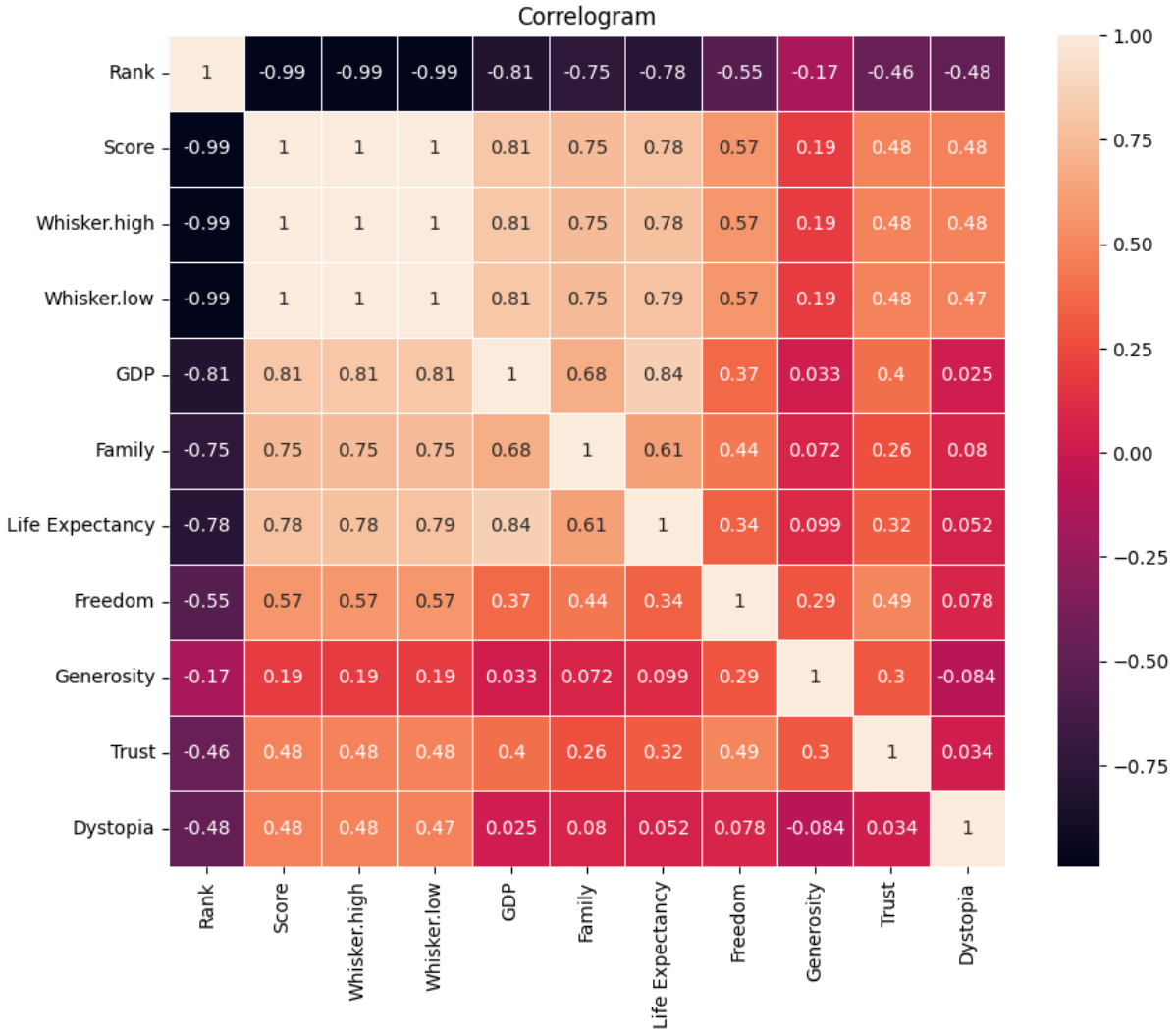
CHƯƠNG III: PHÂN TÍCH TỔNG QUAN

3.1. Trực quan tương quan

Biểu đồ 3.1. Biểu đồ phân phối cho các yếu tố

Để hình dung về mối quan hệ theo cặp giữa các biến các giá trị dựa trên biến Score, nhóm chọn biểu đồ phân tán ma trận để có thể trực quan được phân phối giữa các cặp biến, mối quan hệ tuyến tính giữa các biến và tương quan giữa các biến

Nhận xét: Có thể thấy, các điểm số từ 1 đến 7 đều có phân phối các biến khác nhau với các khoảng giá trị khác nhau. Các biến có tương quan mạnh là ‘GDP - Life Expectancy’, ‘Generosity - Life Expectancy’ , ‘Freedom - Life Expectancy’, ‘GDP - Life Expectancy’



Biểu đồ 3.2. Heatmap – tương quan các biến

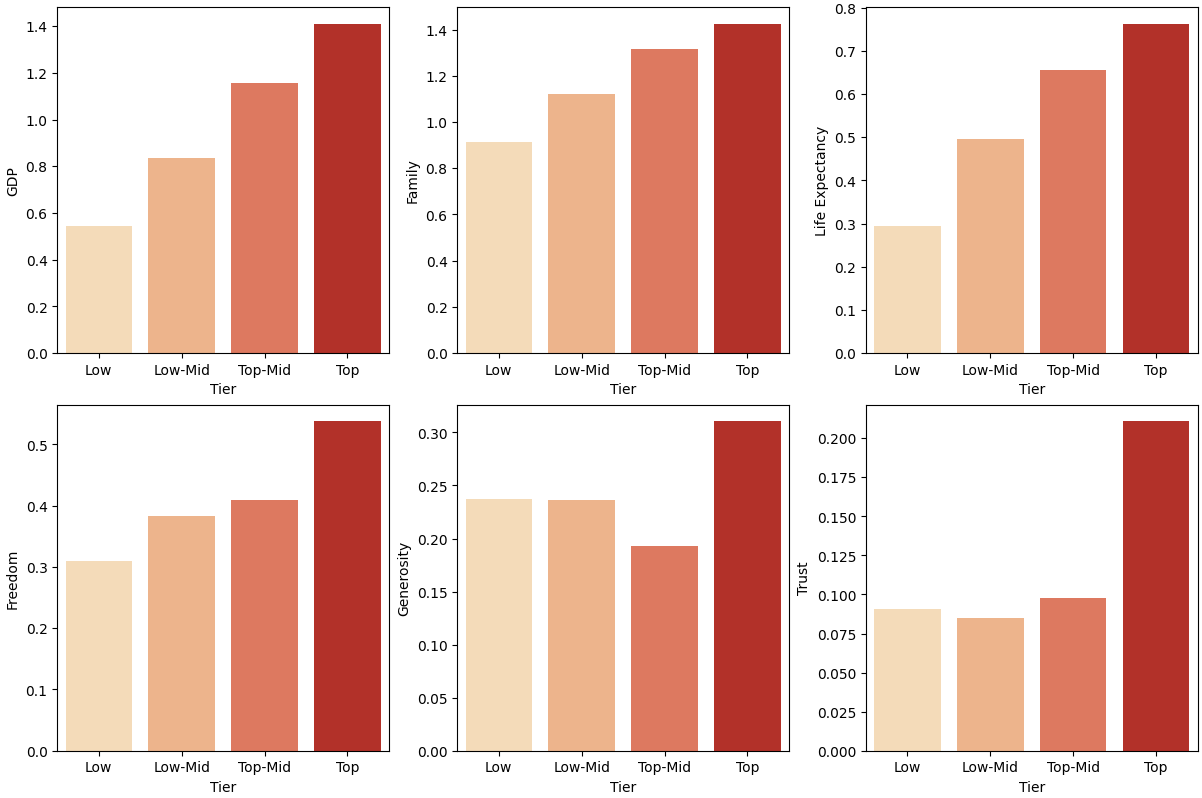
Dựa vào biểu đồ tương quan, màu sắc sáng đại diện cho độ tương quan dương cao, màu sắc càng tối thể hiện cho độ tương quan nghịch cao. Biến Whisker.high và Whisker.low có độ tương đồng lớn với biến Score, với hệ số tương quan đều bằng 1→ 2 biến Whisker.high và Whisker.low đều có ảnh hưởng trực tiếp lên biến Score.

3.2. Trực quan bộ dữ liệu

Với bộ dữ liệu ‘World Score Happiness 2017’ là một bộ dữ liệu xếp hạng từ cao xuống thấp theo thang điểm hạnh phúc, để nhìn rõ hơn về các quốc gia đang nằm ở mức ‘Cao - Trung bình cao - Trung bình - Thấp’, chia theo tứ phân vị 25% để có thể dễ dàng nắm được thông tin hơn khi nhìn vào bộ dữ liệu

	Country	Rank	Score	Whisker.high	Whisker.low	GDP	Family	Life Expectancy	Freedom	Generosity	Trust	Dystopia	Tier
0	Norway	1	7.537000	7.594445	7.479556	1.616463	1.533524	0.796667	0.635423	0.362012	0.315964	2.277027	Top
1	Denmark	2	7.522000	7.581728	7.462272	1.482383	1.551122	0.792566	0.626007	0.355280	0.400770	2.313707	Top
2	Iceland	3	7.504000	7.622030	7.385970	1.480633	1.610574	0.833552	0.627163	0.475540	0.153527	2.322715	Top
3	Switzerland	4	7.494000	7.561772	7.426227	1.564980	1.516912	0.858131	0.620071	0.290549	0.367007	2.276716	Top
4	Finland	5	7.469000	7.527542	7.410458	1.443572	1.540247	0.809158	0.617951	0.245483	0.382612	2.430182	Top
39	Slovakia	40	6.098000	6.177348	6.018652	1.325394	1.505059	0.712733	0.295817	0.136544	0.024211	2.097777	Top-Mid
40	Bahrain	41	6.087000	6.178989	5.995011	1.488412	1.323110	0.653133	0.536747	0.172668	0.257042	1.656149	Top-Mid
41	Malaysia	42	6.084000	6.179980	5.988021	1.291215	1.284646	0.618784	0.402265	0.416609	0.065601	2.004449	Top-Mid
42	Nicaragua	43	6.071000	6.186584	5.955417	0.737299	1.287216	0.653096	0.447552	0.301674	0.130688	2.513931	Top-Mid
43	Ecuador	44	6.008000	6.105848	5.910152	1.000820	1.286169	0.685636	0.455198	0.150112	0.140135	2.290353	Top-Mid
44	El Salvador	45	6.003000	6.108635	5.897364	0.909784	1.182125	0.596019	0.432453	0.078258	0.089981	2.714594	Top-Mid
77	Kosovo	78	5.279000	5.364848	5.193152	0.951484	1.137854	0.541452	0.260288	0.319931	0.057472	2.010541	Low-Mid
78	China	79	5.273000	5.319278	5.226721	1.081166	1.160837	0.741416	0.472788	0.028807	0.022794	1.764939	Low-Mid
79	Pakistan	80	5.269000	5.359984	5.178016	0.726884	0.672691	0.402048	0.235215	0.315446	0.124348	2.792489	Low-Mid
80	Indonesia	81	5.262000	5.352889	5.171112	0.995539	1.274445	0.492346	0.443323	0.611705	0.015317	1.429477	Low-Mid
81	Venezuela	82	5.250000	5.370032	5.129968	1.128431	1.431338	0.617144	0.153997	0.065020	0.064491	1.789464	Low-Mid
150	Rwanda	151	3.471000	3.543030	3.398970	0.368746	0.945707	0.326425	0.581844	0.252756	0.123506	0.540061	Low
151	Syria	152	3.462000	3.663669	3.260331	0.777153	0.396103	0.500533	0.081539	0.493664	0.151347	1.061574	Low
152	Tanzania	153	3.349000	3.461430	3.236570	0.511136	1.041990	0.364509	0.390018	0.354256	0.066035	0.621130	Low
153	Burundi	154	2.905000	3.074690	2.735310	0.091623	0.629794	0.151611	0.059901	0.204435	0.084148	1.683024	Low
154	Central African Republic	155	2.693000	2.864884	2.521116	0.091623	0.629794	0.018773	0.270842	0.280876	0.056565	2.066005	Low

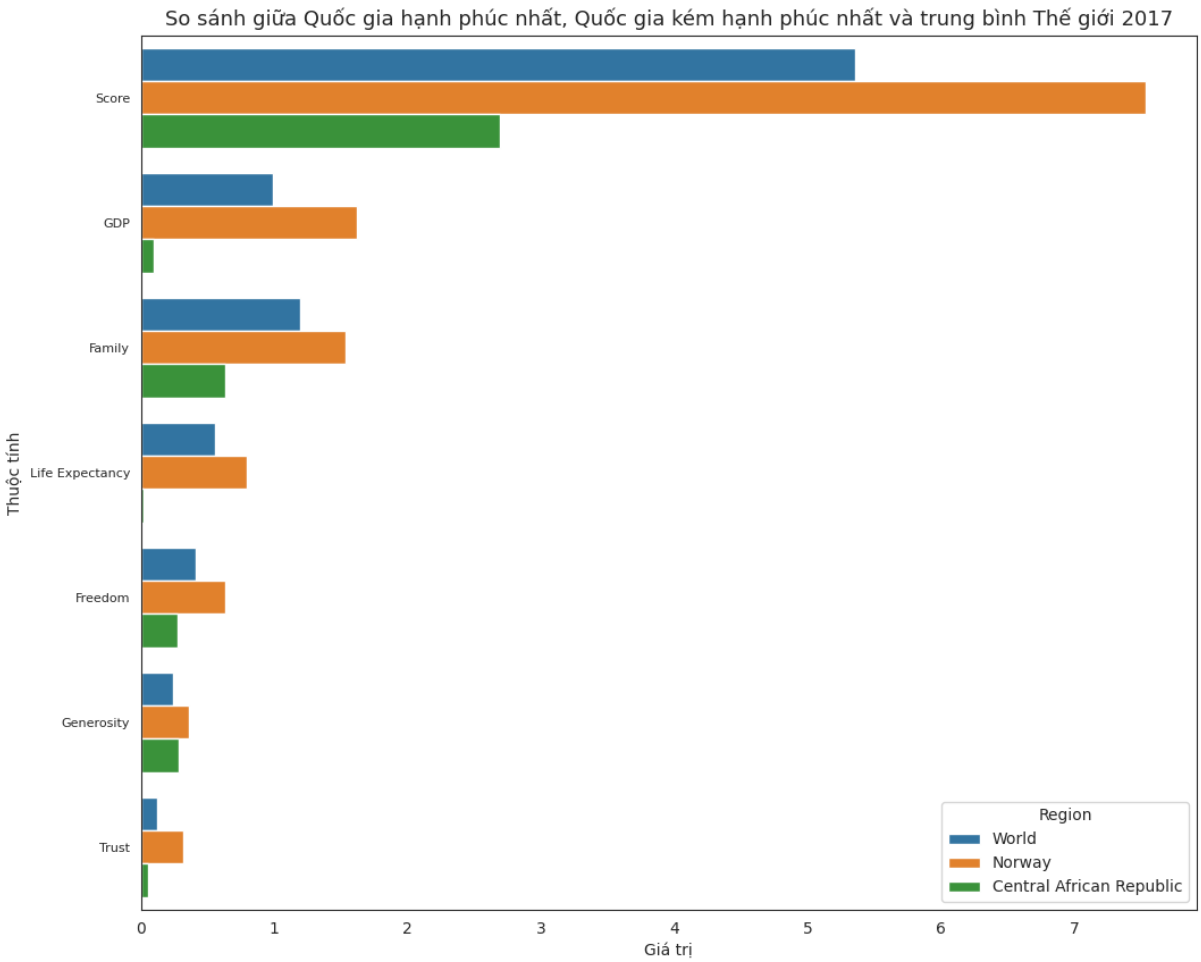
Vì Chỉ số hạnh phúc được đánh giá dựa trên nhiều yếu tố, để phân ra các nước trong bộ dữ liệu nằm ở mức ‘Cao’ - ‘Trung bình - Cao’ - ‘Trung bình - Thấp’- ‘Thấp’ thì cần đánh giá dựa trên các yếu tố khác như: GDP, Family, Life Expectancy, Freedom, Generosity và Trust. Để có thể so sánh các số liệu giữa các nhóm này, nhóm sử dụng biểu đồ cột để so sánh



Biểu đồ 3.3. Barchart - Thể hiện phân phối các biến của 4 nhóm 'Cao' - 'Trung bình - Cao' - 'Trung bình - Thấp'- 'Thấp'

Nhận xét: Đối với GDP và Life Expectancy, có phân phối tương đồng nhau giữa 4 nhóm 'Cao' - 'Trung bình - Cao' - 'Trung bình - Thấp'- 'Thấp'. Tuy nhiên, ở biểu đồ cột biến Trust - 'Mức độ tin tưởng và sự nhận thức về tham nhũng trong chính phủ' của các nước thuộc nhóm Trung bình - Thấp, thấp hơn một nửa so với các nước có Chỉ số hạnh phúc thuộc mức 'Cao'.

Dựa trên bộ dữ liệu sau khi đã phân các nước theo 4 nhóm dựa trên Chỉ số hạnh phúc, quốc gia có điểm số hạnh phúc cao nhất là Norway và quốc gia có số điểm hạnh phúc thấp nhất là Central African Republic. Để thấy được sự chênh lệch giữa điểm hạnh phúc của trung bình các nước với 2 nước có điểm hạnh phúc cao nhất và thấp nhất trong bộ dữ liệu, nhóm biểu đồ cột để biểu diễn điều này trực quan hơn.



Biểu đồ 3.4. Bar chart - So sánh giữa quốc gia hạnh phúc nhất, quốc gia kém hạnh phúc nhất và trung bình thế giới 2017

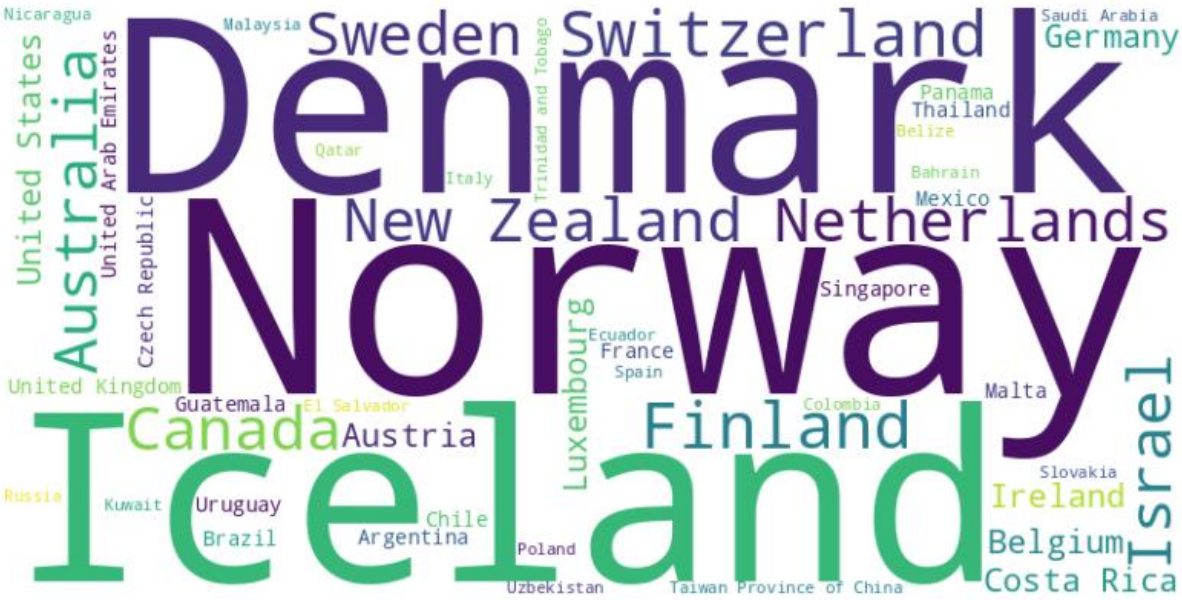
Nhận xét: Dựa trên biểu đồ, Score - Điểm số hạnh phúc của Central African Republic thấp hơn một nửa so với trung bình thế giới và chỉ bằng khoảng ⅓ so với nước hạnh phúc nhất - Norway. Norway luôn có các chỉ số cao hơn so với trung bình thế giới. Đối với các chỉ số như GDP, Life Expectancy, Trust, Central African Republic thấp hơn rất nhiều so với trung bình các nước trên thế giới, và đặc biệt, Life Expectancy gần như là bằng 0. Tuy nhiên, đối với Generosity - tỷ lệ quyên góp ở quốc gia này lại cao hơn so với trung bình.

Tóm lại có thể thấy, Quốc gia hạnh phúc nhất có chỉ số cao nhất ở GDP, Family, Life Expectancy và Freedom, cho thấy những yếu tố này có ảnh hưởng lớn đến hạnh phúc của người dân. Ngược lại, với quốc gia kém hạnh phúc nhất có chỉ số thấp nhất ở hầu hết các thuộc tính, ngoại trừ Generosity, đã gây ra nhiều khó khăn và bất công cho người dân, làm giảm hạnh phúc của họ.

3.3. Trực quan tỷ lệ

Lý do cho việc chọn 50 quốc gia có xếp hạng cao nhất là vì đây là những quốc gia lớn mạnh và có tầm ảnh hưởng trên thế giới. Điều này giúp ta có cái nhìn tổng thể về mức độ hạnh phúc của các quốc gia đó.

Việc trực quan 50 quốc gia là số lượng lớn, các dạng biểu đồ khác có thể sẽ bị quá tải thông tin và trở nên khó nhìn. Do vậy ở đây ta chọn dạng Word Cloud vì muốn tạo ra hình ảnh trực quan, ngắn gọn và thu hút người xem.



Biểu đồ 3.5. Word Cloud - BIểu đồ thể hiện top 50 quốc gia hạnh phúc nhất thế giới

Nhận xét: Biểu đồ cho thấy sự phân bố của các quốc gia dựa trên điểm hạnh phúc (Score). Các quốc gia có chỉ số hạnh phúc cao xuất hiện ở trung tâm và kích thước chữ lớn hơn hẳn và ngược lại. Điều này giúp người đọc dễ dàng tiếp nhận thông tin kể cả có kiến thức về phân tích dữ liệu hay không. Tuy nhiên biểu đồ này có hạn chế là gây khó khăn cho việc so sánh dữ liệu, đặc biệt là những quốc gia có Score thấp.

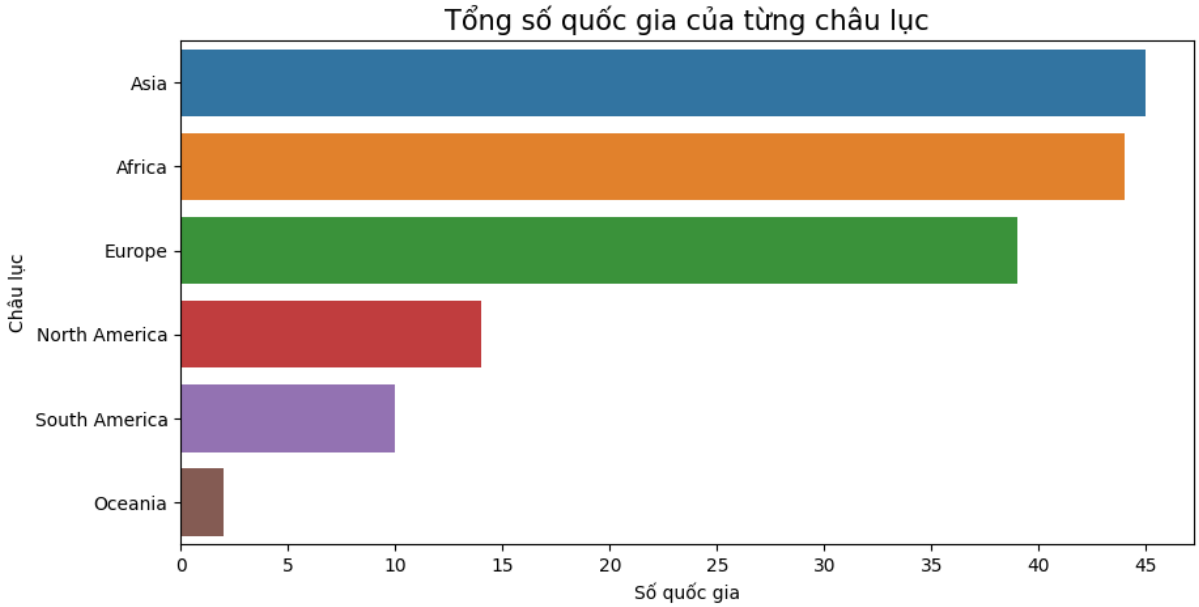
CHƯƠNG IV: PHÂN TÍCH THEO CHÂU LỤC

4.1. Trực quan tổng thể

4.1.1. Theo tổng thể

Bộ dữ liệu "World Happiness Report - 2017" có tận 155 quốc gia, điều này rất khó để biểu diễn dữ liệu trong cùng một lúc mà vẫn mang đến thông tin dễ nhìn và ngắn gọn. Vì thế ta có thể gom nhóm chúng theo từng khu vực địa lý, việc này giúp chúng ta nhìn nhận rõ hơn về xu hướng và đặc điểm riêng của từng châu lục.

Trong trường hợp này, ta cần dùng biểu đồ phù hợp biểu diễn dữ liệu theo tỷ lệ, và đó là Bar chart. Vì biểu đồ này có thể biểu diễn chiều cao của cột tương ứng.



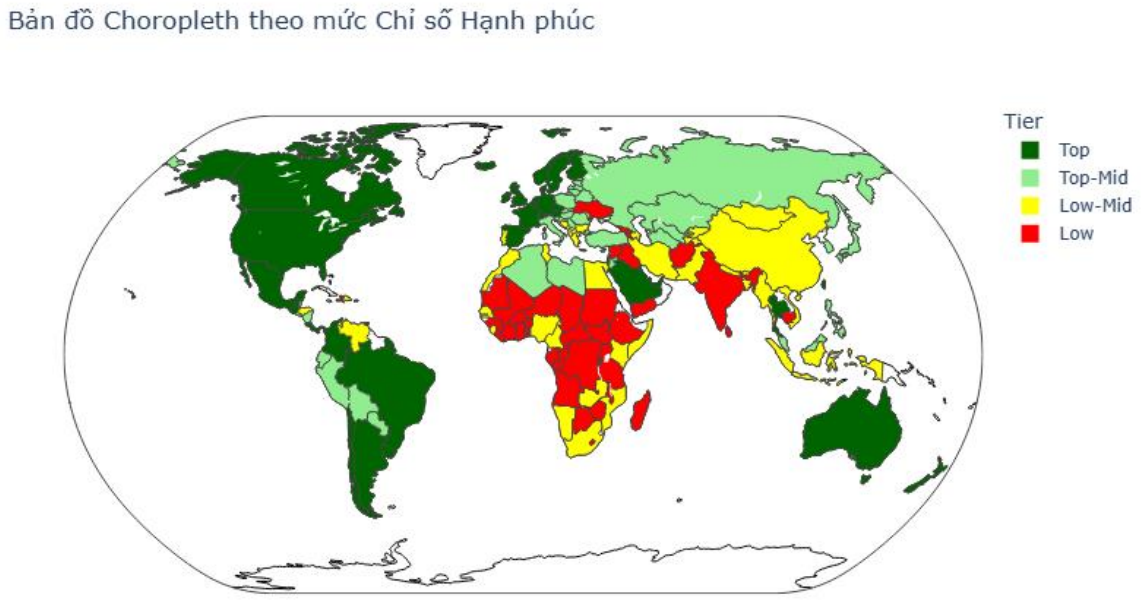
Biểu đồ 4.1. Bar chart - Tổng số quốc gia của từng khu vực

Nhận xét: Biểu đồ cho thấy Châu Á là khu vực có số quốc gia nhiều nhất trên thế giới, với 45 quốc gia. Châu Phi đứng thứ hai với 44 quốc gia. Châu Âu đứng thứ ba với 39 quốc gia. Các vị trí tiếp theo lần lượt là Bắc Mỹ với 14 quốc gia, Nam Mỹ với 10 quốc gia và Châu Đại Dương chỉ 2 quốc gia.

4.1.2. Theo nhóm chỉ số Hạnh phúc

Để dễ dàng hình dung mức độ hạnh phúc của một quốc gia so với toàn cầu, nhóm chia các nước theo tứ phân vị 25%, mỗi nước sẽ thuộc một trong 4 nhóm dựa trên chỉ số hạnh phúc, và sẽ ứng với các châu lục khác nhau. Tiếp theo, nhóm kiểm tra các quốc gia thuộc mức Chỉ số hạnh phúc nào trong nhóm Top, Top-Mid, Low-Mid và Low.

Vì Tier là biến định tính, biểu đồ được trực quan hóa sẽ gọn gàng và dễ hình dung hơn so với khi ta sử dụng Score (biến định tính) cho biểu đồ. Để thực hiện công việc này, ta chọn biểu đồ Choropleth để thể hiện dữ liệu địa lý, cùng với biến Country và Tier.



Biểu đồ 4.2. Choropleth - Phân phối Tier theo Country

Nhận xét:

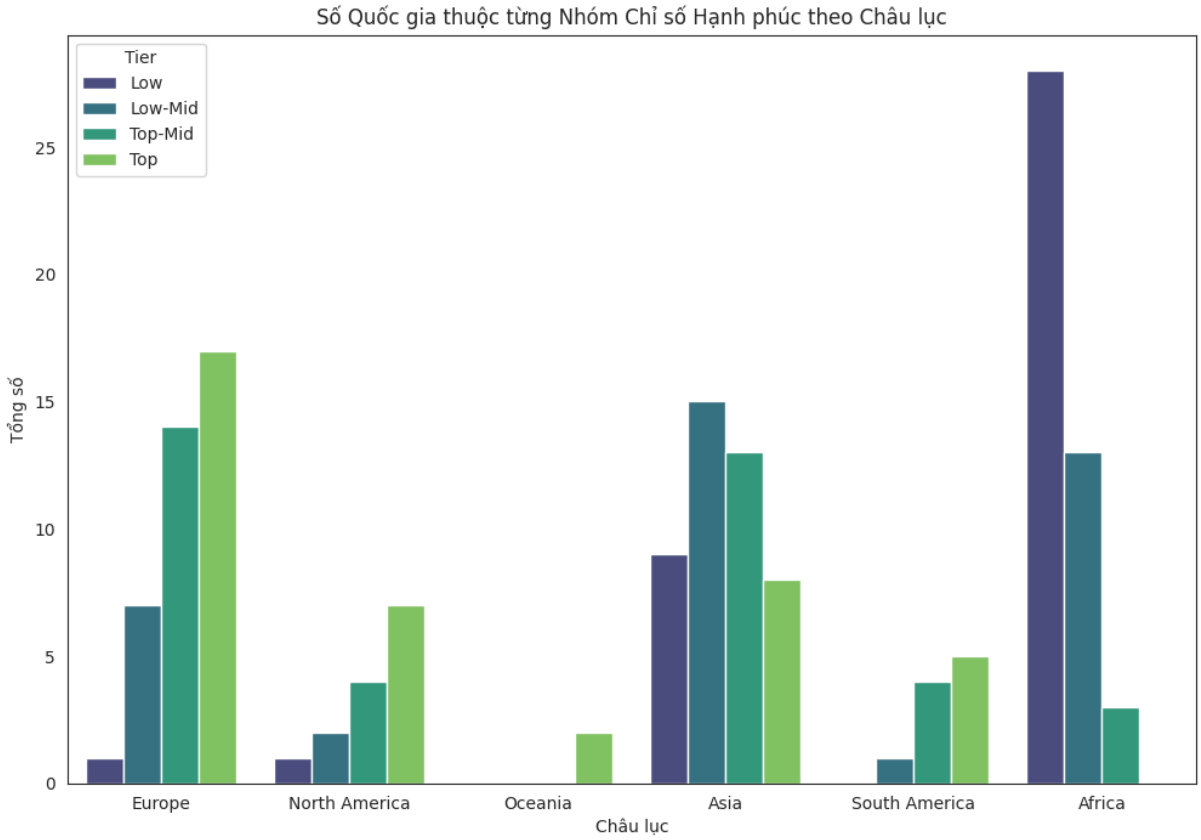
Có thể thấy, khu vực Châu Mỹ có số lượng các quốc gia trong nhóm có Chỉ số Hạnh phúc cao. Phần lớn các quốc gia đều thuộc nhóm Top, một số quốc gia Nam Mỹ thuộc nhóm Top-Mid hoặc Low-Mid. Chỉ có Haiti là đất nước thuộc nhóm Low.

Châu lục nhìn chung có Chỉ số Hạnh phúc tốt tiếp theo là Châu Âu, với nhiều quốc gia thuộc nhóm Top hoặc Top-Mid.

Tiếp tục, Châu Á có phần lớn các quốc gia thuộc nhóm Low-Mid và Top-Mid, tuy nhiên, khác với các châu lục khác, Châu Á có các quốc gia thuộc mọi Tier, và phân phối của các Tier có vẻ đều hơn các châu lục khác.

Mặt khác, khu vực Châu Phi lại chỉ gồm các quốc gia có Chỉ số Hạnh phúc nằm dưới Phân vị 2 của dữ liệu (nhóm Low Mid & Low). Ngoại lệ là Algeria và Libya, với Chỉ số Hạnh phúc nằm giữa Phân vị 2 và 3 (nhóm High-Mid).

Để kiểm tra quan sát trên rằng Châu Á có phân phối Tier đều hơn các Châu lục khác, ta tiến hành vẽ biểu đồ để hiểu rõ hơn về tổng số quốc gia thuộc một Tier theo từng châu lục. Để thể hiện tổng số theo nhóm, ta chọn biểu đồ cột.



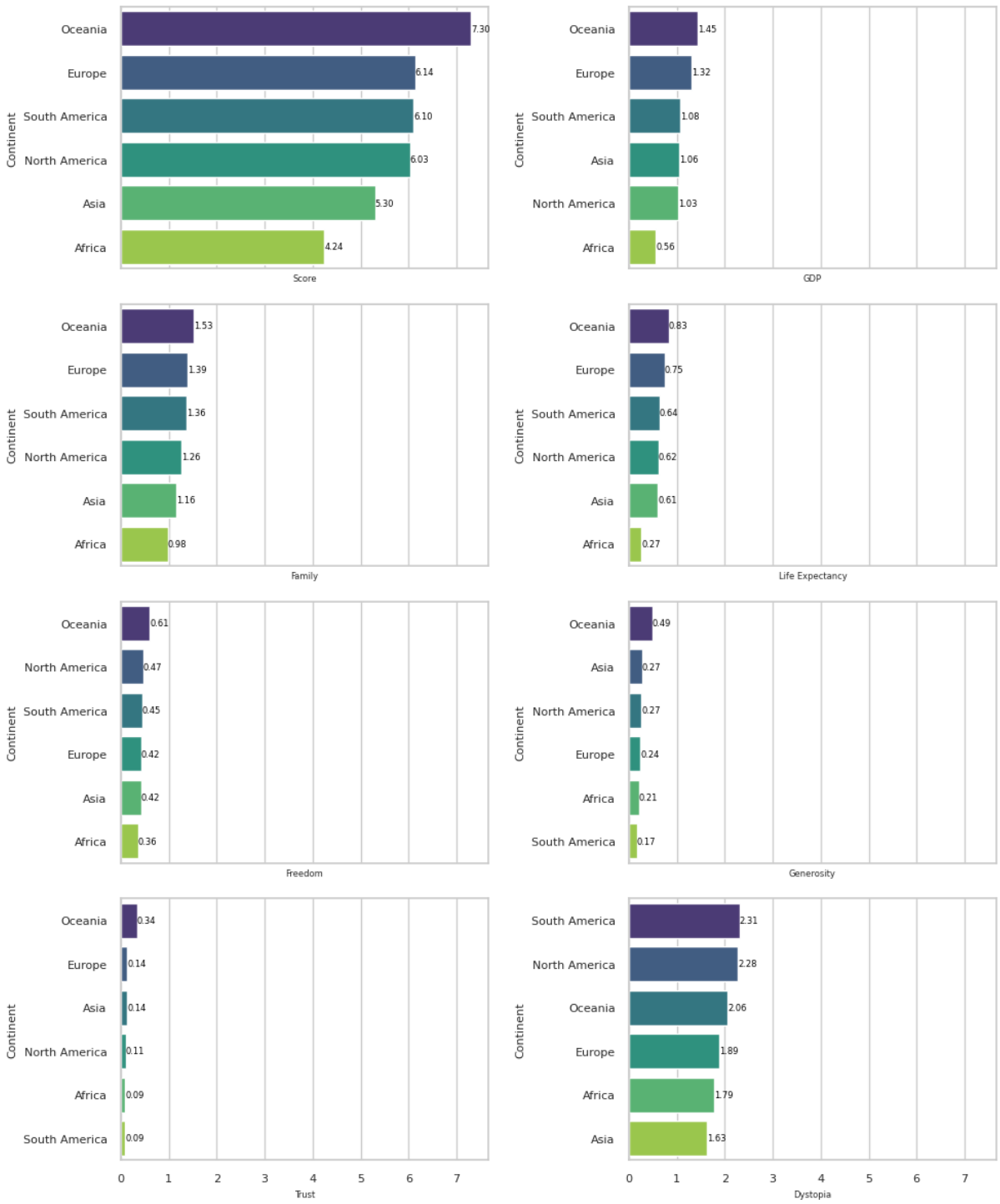
Biểu đồ 4.3. Bar graph - Tổng số Country theo Tier, nhóm theo Continent

Nhận xét:

Đúng như quan sát, Châu Á có phân phối ít lệch hơn các châu lục khác. Châu Phi có nhiều quốc gia nằm ở nhóm kém hạnh phúc nhất, cũng như Châu Âu có số quốc gia trong nhóm hạnh phúc nhất cao nhất.

4.1.3. Theo giá trị trung bình

Về tổng thể các yếu tố thuộc từng châu lục, để hiểu rõ hơn về độ lớn của các giá trị ở từng châu lục so với toàn cầu, nhóm tính trung bình các giá trị dựa theo châu lục và trình bày ở dạng biểu đồ cột.

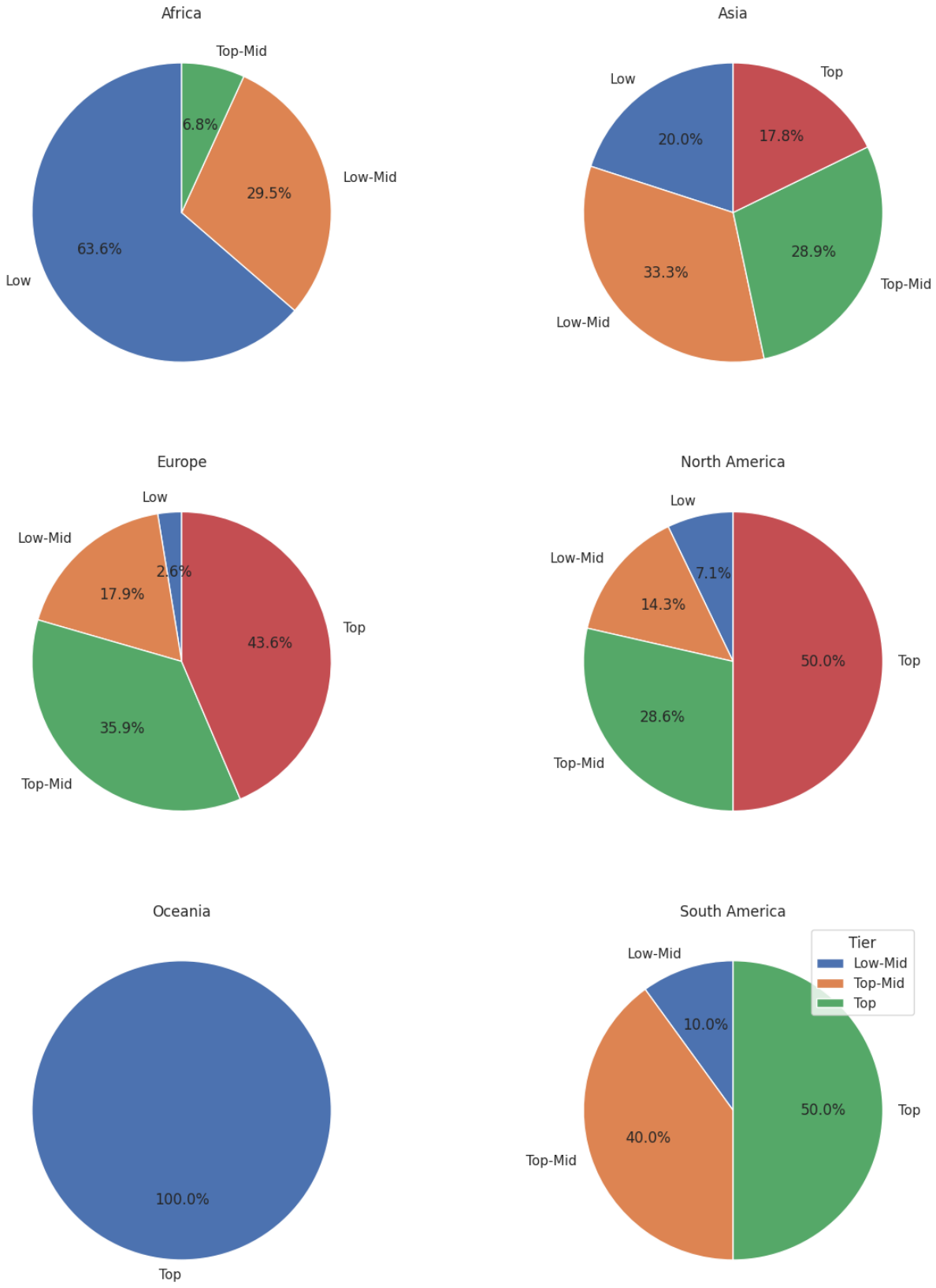


Biểu đồ 4.4. Bar graph - Trung bình các yếu tố, nhóm theo Continent

Nhận xét: Châu Âu là châu lục có chỉ số Hạnh phúc trung bình cao nhất với 6.14 (khi không bao gồm Châu Đại dương). Ngoài ra, Châu Âu cũng có giá trị trung bình của 4/8 biến cao hơn các châu lục còn lại, cụ thể là các biến Score, Trust, Family, Life Expectancy và GDP. Ngoài ra, Bắc Mỹ và Nam Mỹ có giá trị trung bình của các biến nằm kề nhau ở Score, Family, Life Expectancy, Dystopia và Freedom, cho thấy sự tương đồng về bối cảnh kinh tế - xã hội của hai phần thuộc Châu Mỹ.

4.2. Trực quan tỷ lệ

Dù biểu đồ cột có thể cho thấy tổng số nhóm chỉ số Hạnh phúc tại mỗi Châu lục, vì mỗi Châu lục có số quốc gia khác nhau, nhóm muốn tìm hiểu tại 6 châu lục thì tỷ lệ % mỗi nhóm nước theo chỉ số hạnh phúc trong từng châu lục sẽ là bao nhiêu. Để trực quan hóa điều này, nhóm sử dụng biểu đồ tròn để thể hiện tỷ lệ mỗi châu lục sẽ có bao nhiêu phần trăm các nước thuộc 4 nhóm điểm.



Biểu đồ 4.5. Pie chart - Tỷ lệ Tier, nhóm theo Continent

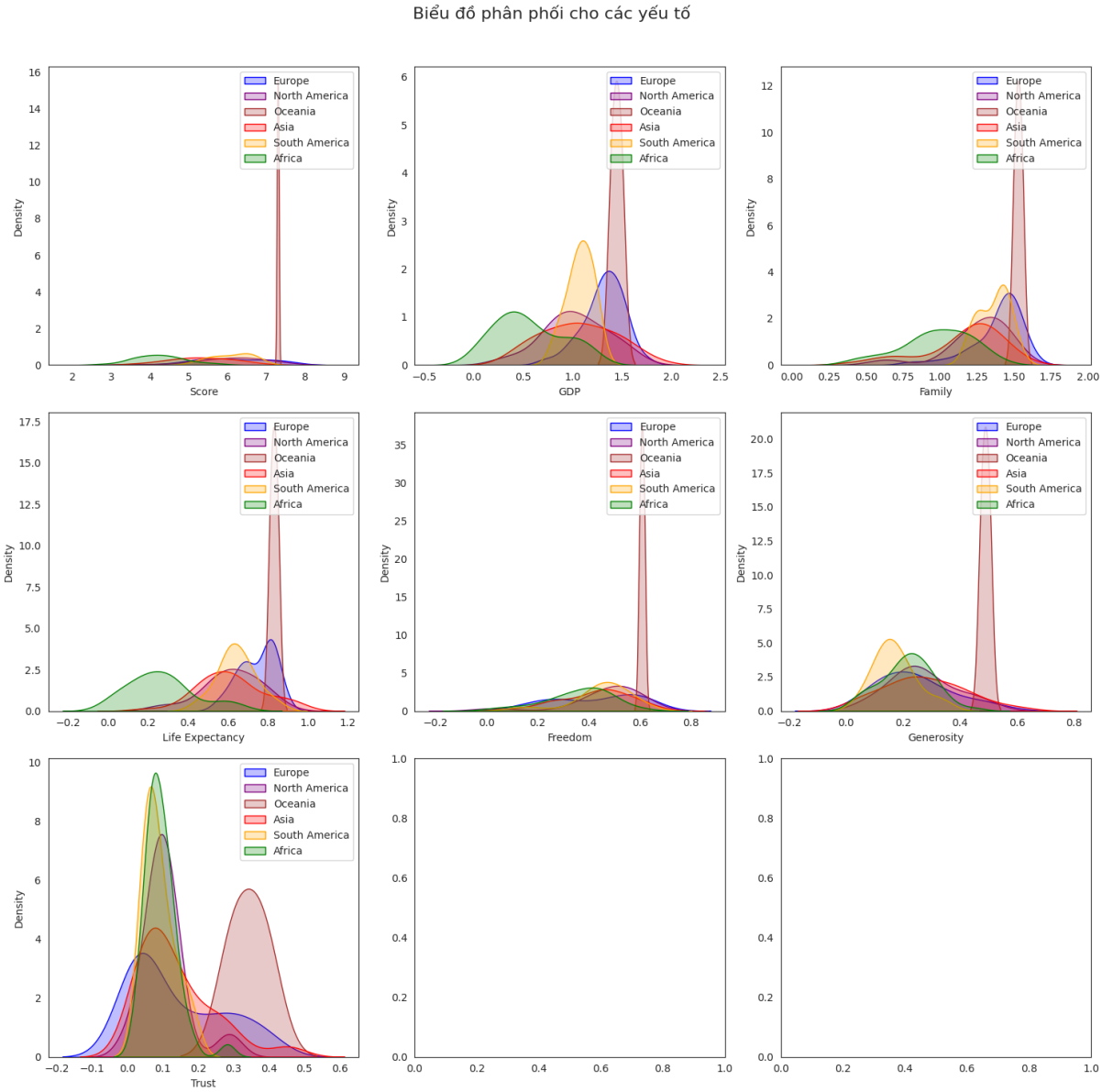
Có vẻ như Châu Phi vẫn có tỷ lệ số quốc gia thuộc nhóm Low cao nhất. Tuy nhiên, lúc này tỷ lệ Top lại cao nhất ở Bắc Mỹ và gần bằng với của Nam Mỹ. Các quốc gia ở Châu Úc đều nằm trong nhóm Top,

Châu Phi chỉ bao gồm các quốc gia có chỉ số hạnh phúc ở mức Thấp, Trung bình - Thấp, và số ít các nước thuộc các nhóm Trung bình - Cao. Châu Á có sự phân bố tương đối đều giữa 4 mức chỉ số hạnh phúc. Châu Âu và Bắc Mỹ có sự tương đồng giữa về sự phân bố các nhóm chỉ số. Bên cạnh đó, Bắc Mỹ có tỷ lệ các nước có chỉ số hạnh phúc ở mức 'Cao' nhiều hơn so với châu Âu, tuy nhiên, Bắc Mỹ có tỷ lệ các nước có chỉ số hạnh phúc 'Thấp' cao hơn so với châu Âu. Châu Đại Dương là châu lục duy nhất trong 6 châu lục có toàn bộ các quốc gia có chỉ số hạnh phúc 'Cao' nhưng vì dữ liệu chỉ gồm hai quốc gia thuộc Châu lục này nên tỷ lệ theo nhóm của châu lục này không đáng kể. Nam Mỹ không có quốc gia nào có chỉ số hạnh phúc 'Thấp'

Tóm lại: 4 Châu lục có tỷ lệ các nước có chỉ số hạnh phúc 'Cao': châu Âu, Bắc Mỹ, Nam Mỹ và châu Đại dương. Châu Phi có nhiều quốc gia thuộc nhóm có chỉ số Hạnh phúc thấp nhất trong 6 châu lục.

4.3. Trực quan phân phối

Để hình dung về phân phối các giá trị ở các Châu lục, nhóm chọn biểu đồ phân tán và biểu đồ phân phối dựa trên từng cặp biến.

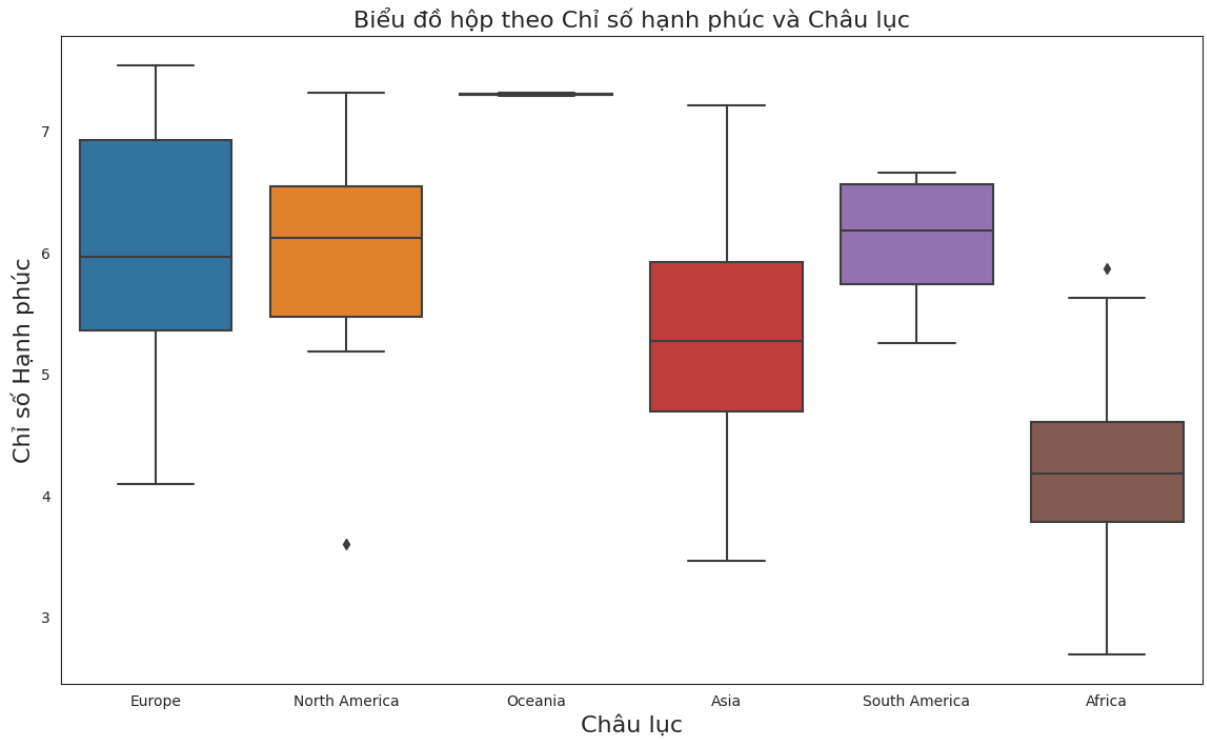


Biểu đồ 4.6. Pair plot - Phân phối & tương quan các cặp biến, nhóm theo Continent

Nhận xét: Có thể thấy, các châu lục đều có phân phối các biến khác nhau với các khoảng giá trị khác nhau. Trong đó, hai biến có khoảng giá trị tương đồng nhất giữa các châu lục là Generosity.

Mặt khác, các biến thuộc Châu Úc có khoảng phân phối ngắn hơn, nhưng đó là vì bộ dữ liệu chỉ bao gồm hai quốc gia thuộc châu lục này.

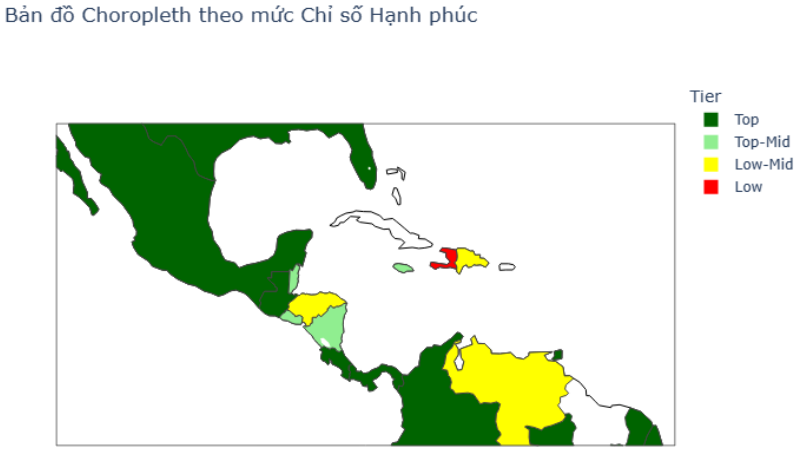
Vì Chỉ số Hạnh phúc là yếu tố chính cần đánh giá, nhóm tiến hành vẽ biểu đồ hộp để nhìn rõ hơn về phân phối của biến này theo nhóm các Châu lục.



Biểu đồ 4.7. Box plot - Phân phối của Score, nhóm theo Continent

Nhận xét: Châu Á là châu lục có khoảng giá trị của chỉ số Hạnh phúc rộng nhất, tiếp đến là Châu Âu. Mặc dù Châu Âu có giá trị trung bình của chỉ số Hạnh phúc cao hơn các châu lục khác, 75% giá trị của châu lục này cũng nằm trong khoảng rộng hơn các châu lục khác. Điều này cho thấy mức độ hạnh phúc của người dân ở châu lục này biến động nhiều hơn theo từng quốc gia.

Ngoài ra, ta có thể thấy điểm dữ liệu ngoại lai ở Châu Mỹ, đất nước có giá trị thấp hơn mọi quốc gia tại châu lục này. Quốc gia này là Haiti (được đánh dấu màu đỏ), là đất nước hàng xóm của Dominican Republic (được đánh dấu màu vàng) nhưng với chỉ số Hạnh phúc thấp hơn đáng kể.



Biểu đồ 4.8. Choropleth - Phân phối Tier theo Country (Haiti)

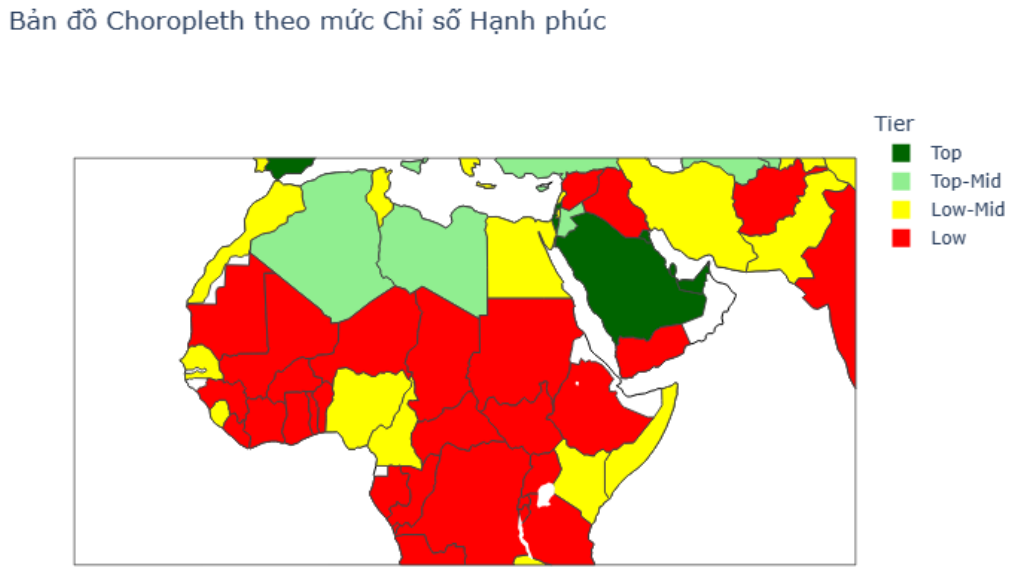
Để hiểu rõ hơn vì sao quốc gia này lại có chỉ số Hạnh phúc thấp hơn, nhóm tiến hành so sánh các giá trị của quốc gia này so với trung bình châu lục của đất nước, với toàn thế giới và với nhóm "Low".

Biểu đồ 4.9. Bar graph - Giá trị trung bình nhóm so với Haiti

Nhận xét: Có thể thấy, Haiti có giá trị thấp hơn mọi nhóm ở phần lớn các biến bao gồm Score, GDP, Family, Life Expectancy và đặc biệt là Freedom.

Cụ thể hơn về chỉ số Freedom của quốc gia này, có thể nhắc đến dữ liệu từ "Freedom in the World". Đây là một bản báo cáo toàn cầu hàng năm về quyền chính trị và tự do dân sự, bao gồm các đánh giá số liệu và mô tả cho mỗi quốc gia và một số lãnh thổ được chọn. "Freedom in the World" hoạt động dựa trên giả định rằng sự tự do đạt được tốt nhất trong các xã hội dân chủ. Theo cơ sở dữ liệu của tổ chức này, vào năm 2017, Haiti được đánh dấu "Partly Free" - tự do bán phần. Theo Freedom House, các quốc gia Partly Free có thể đủ tiêu chuẩn là các nền dân chủ bầu cử (electoral), nhưng không tự do (liberal). Tuy nhiên, cũng theo tổ chức này, vào năm 2023, Haiti đã giảm từ Partly Free xuống Not Free (không tự do) do sự kiện ám sát Tổng thống Jovenel Moïse, sự sụp đổ liên tục trong hệ thống bầu cử và các cơ quan nhà nước khác, cũng như ảnh hưởng xấu của tội phạm tổ chức và bạo lực đối với cuộc sống công dân.

Hai điểm dữ liệu ngoại lai khác ở Châu Mỹ thuộc về đất nước Algeria và Libya. Đây là hai quốc gia ven biển thuộc phía Bắc Châu Phi, cũng là hai quốc gia duy nhất ở châu lục này nằm trong nhóm "Top-Mid".



Biểu đồ 4.10. Choropleth - Phân phối Tier theo Country (Algeria & Libya)

Tương tự, nhóm thực hiện trực quan hóa giá trị trung bình giữa các nhóm khác nhau và so sánh với hai quốc gia này.

Biểu đồ 4.11Bar graph - Giá trị trung bình nhóm so với Haiti (Algeria & Libya)

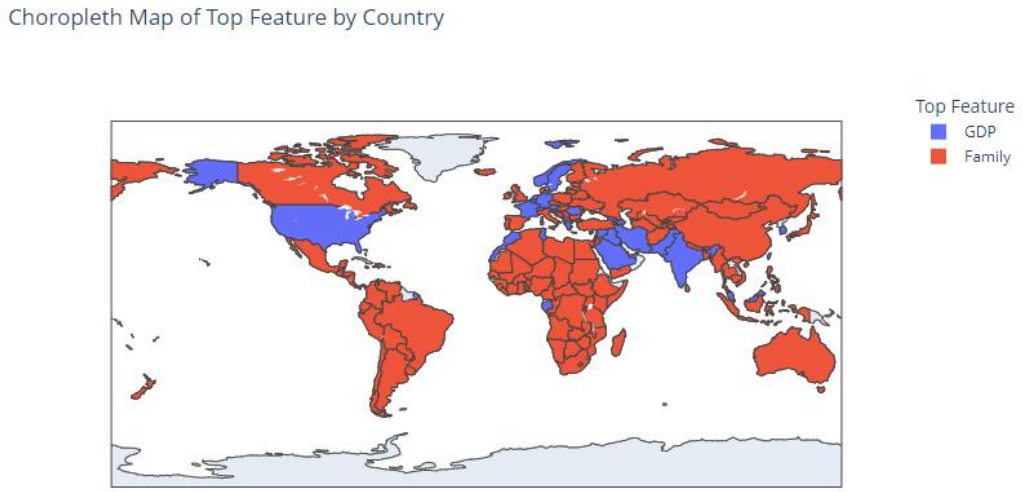
Nhận xét:

Algeria và Libya có GDP cao hơn trung bình của thế giới và của toàn châu lục. Đây có thể là vì đây là hai quốc gia ven biển, tạo điều kiện cho việc giao thương với thế giới. Ngoài ra, Libya cũng có chỉ số về Family, Freedom cao hơn trung bình toàn cầu.

Mặt khác, Algeria có Trust, Dystopia và Life Expectancy cao hơn trung bình toàn cầu. Đáng chú ý thay, quốc gia này (Algeria) có chỉ số về tự do và sự hào phóng thấp hơn trung bình Châu Phi. Cũng theo Freedom House, quốc gia này thuộc nhóm "Not Free" khi nói về độ tự do. Đây có thể là vì Algeria có nhiều đảng đối lập, gian lận trong bầu cử với các vấn đề khác bao gồm đàn áp các cuộc biểu tình đường phố, hạn chế pháp lý đối với tự do truyền thông và tham nhũng nặng nề (theo Freedom House).

4.4. Trực quan tương quan

“Độ quan trọng của từng yếu tố kinh tế - xã hội đến chỉ số Hạnh phúc có khác nhau ở từng quốc gia/ châu lục không?”. Để trả lời câu hỏi này, nhóm sẽ sử dụng hệ số tương quan Pearson Correlation để tính tương quan giữa các cặp số theo từng Châu lục. Sau đó, biến có giá trị tương quan cao nhất với chỉ số Hạnh phúc sẽ được dùng để tô màu cho quốc gia trên bản đồ Choropleth.



Biểu đồ 4.12. Choropleth - Phân phối Tier theo Country (Algeria & Libya)

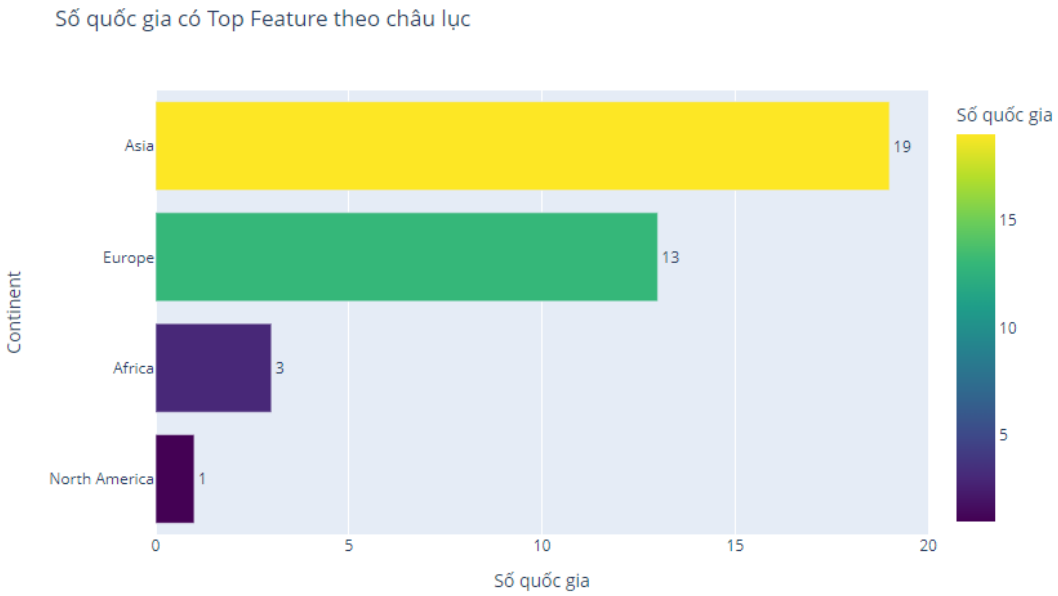
Nhận xét: Hai yếu tố có tương quan cao nhất với Score cho các quốc gia bao gồm có Family và GDP, đặc biệt là yếu tố đầu tiên. Để hiểu rõ hơn về số lượng thật sự, nhóm minh họa số quốc gia có từng loại. Top Feature bằng biểu đồ cột

Biểu đồ 4.13 Choropleth - Phân phối Tier theo Country (Algeria & Libya)

Nhận xét:

Có thể thấy, yếu tố gia đình có tương quan cao nhất với chỉ số Hạnh phúc ở 118 quốc gia. Trong phần lớn các quốc gia, nếu một người có sự hỗ trợ từ người thân những lúc khó khăn, người đó có khả năng

Với 36 quốc gia còn lại trong bộ dữ liệu, GDP là yếu tố có tương quan cao nhất.



Biểu đồ 1Biểu đồ 4. 44. Choropleth - Phân phối Tier theo Country (Algeria & Libya)

CHƯƠNG V: MACHINE LEARNING

5.1. Phân Cụm Quốc Gia dựa trên Score và GDP :

5.1.1. Mục Đích:

Bộ dữ liệu chứa thông tin về chỉ số hạnh phúc (Score), sản lượng kinh tế (GDP), và quốc gia tương ứng. Đối với 155 quốc gia khác nhau, việc phân cụm dựa trên các yếu tố này có thể giúp chúng ta nhận biết nhóm quốc gia có đặc điểm tương đồng về mức độ hạnh phúc và sức mạnh kinh tế.

5.1.2. Phương Pháp:

Chúng ta sẽ sử dụng thuật toán K-means để phân chia quốc gia thành các nhóm dựa trên điểm số hạnh phúc và GDP. Việc này giúp tạo ra các cụm quốc gia có đặc điểm chung, giúp chúng ta hiểu rõ hơn về sự liên quan giữa mức độ hạnh phúc và phương tiện kinh tế.

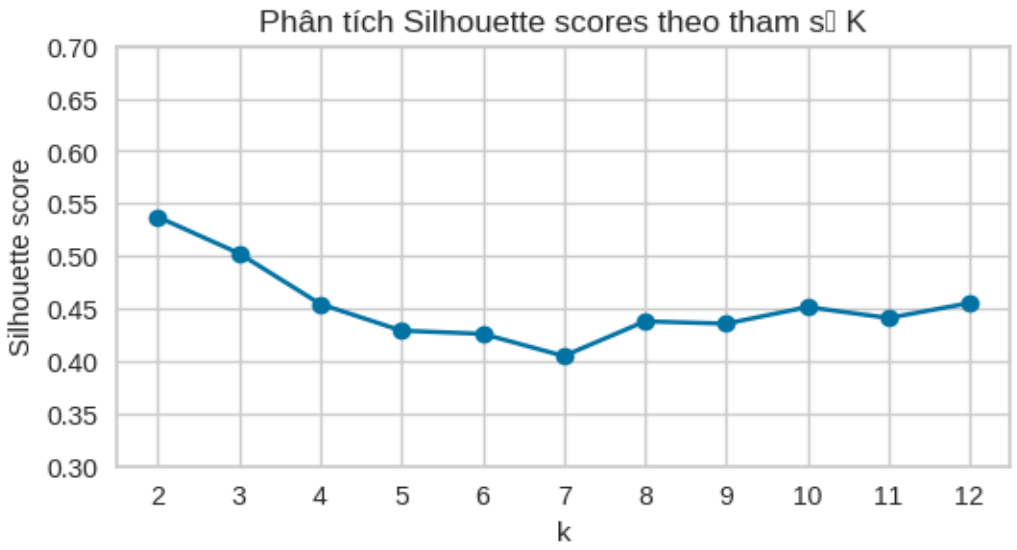
5.1.3. Lợi Ích:

Phân cụm quốc gia giúp chúng ta nhận biết các nhóm có đặc điểm chung về mức sống và phát triển kinh tế. Thông qua việc trực quan hóa dữ liệu, chúng ta có thể dễ dàng so sánh mức độ hạnh phúc và GDP giữa các nhóm, giúp xác định xu hướng và sự tương quan.

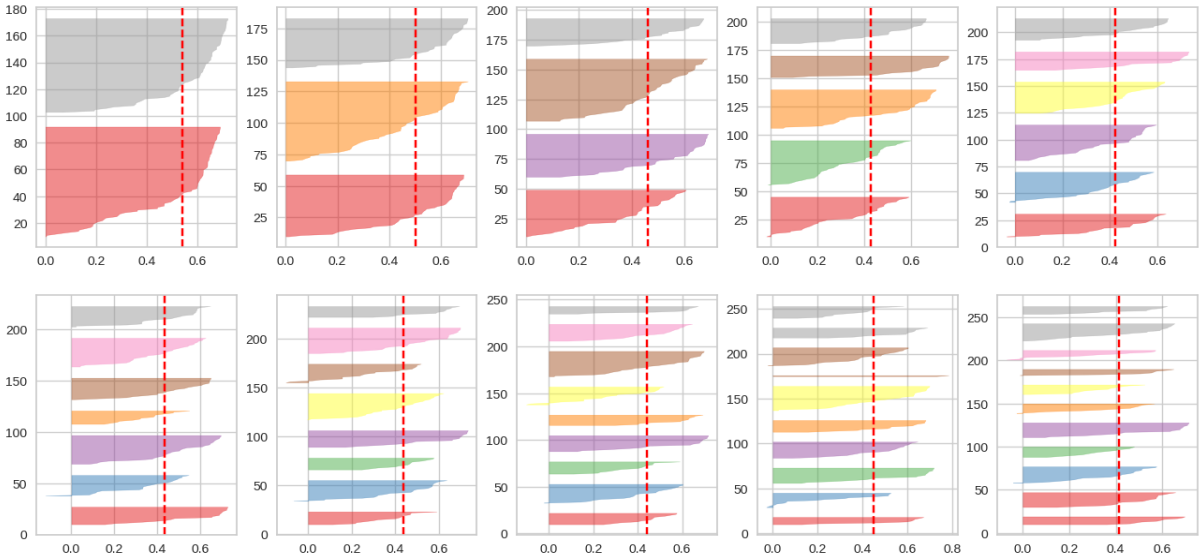
5.1.4. Phân cụm với K-means:

Đầu tiên sử dụng phương pháp elbow để tìm được số cụm tối ưu nhất. Phương pháp elbow sử dụng độ tương quan nghịch giữa SSE (Sum of Square Error) và số cụm. Ở đây, ta sử dụng biểu đồ đường để có thể dễ dàng nhìn thấy sự tương quan này, Và theo biểu đồ, 2 là số cụm tối ưu.

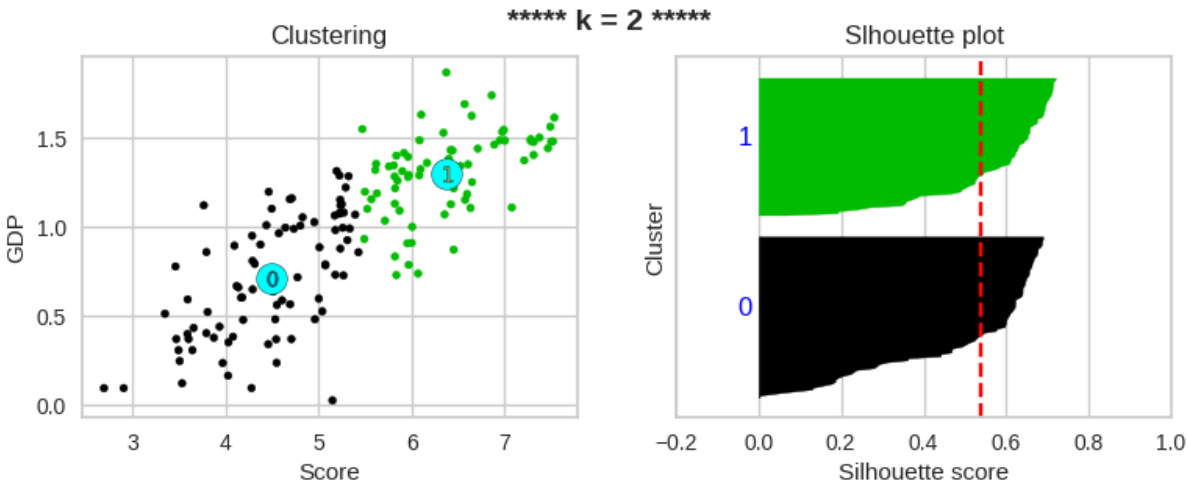
Dùng 3 biểu đồ để thể hiện số cụm tối ưu nhất:



Biểu đồ 5.1. Phân tích Silhouette scores theo giá trị của tham số k



Biểu đồ 5.2. Biểu diễn trực quan Silhouette plot bằng thư viện YellowBrick



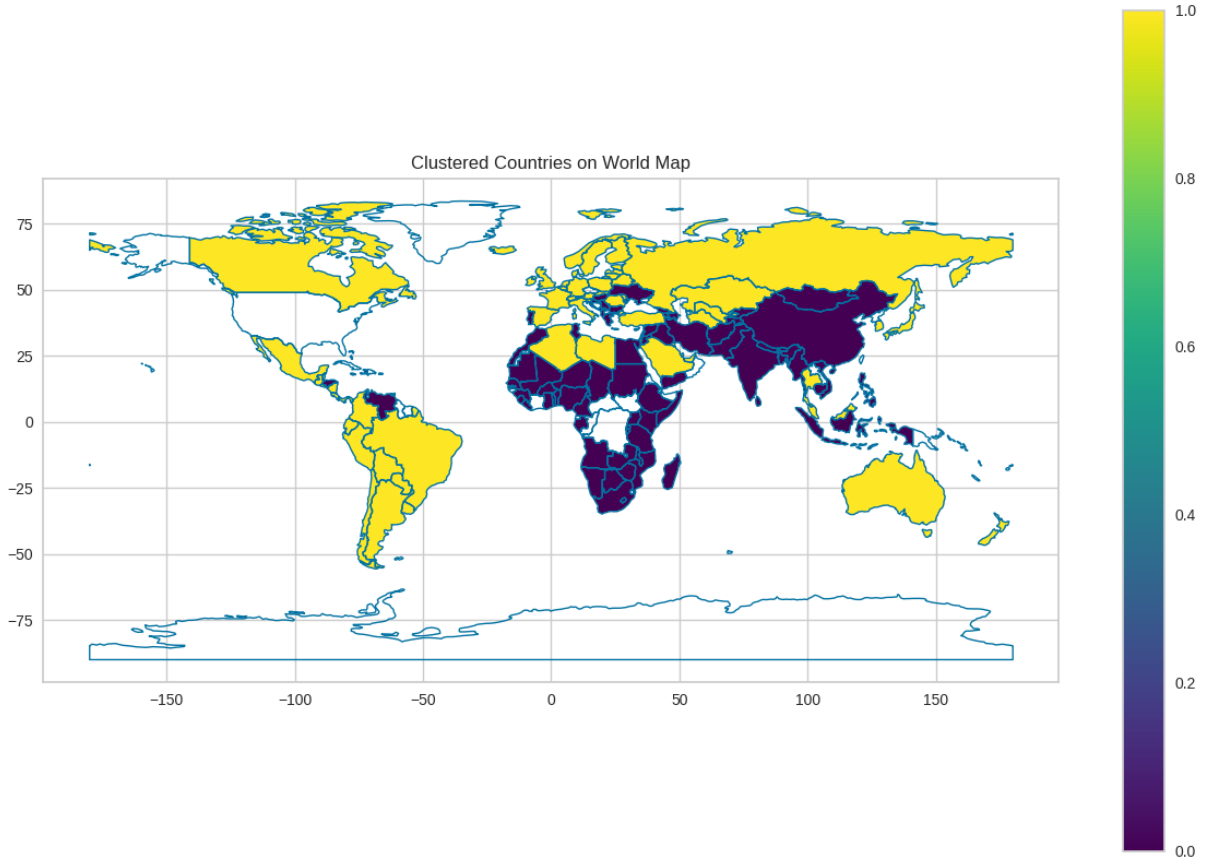
Biểu đồ 5.3. Biểu diễn trực quan clustering và phân tích silhouette scores theo tham số k

5.1.5. Đánh giá biểu đồ

Ba biểu đồ đã dùng ở phần này thể hiện tốt các thông tin cần truyền tải đến người xem, màu sắc dễ nhìn.

Đặc biệt là line chart có thể nhấp vào được từng ô dữ liệu để xem kĩ các phần dữ liệu được biểu diễn

Sau khi đã chọn được hai cụm, ta sẽ thể hiện các quốc gia cùng 1 cụm lại với nhau, ta nhận xét các quốc gia gần nhau sẽ cùng nằm 1 cụm (cùng châu lục). Các châu lục như Châu Á, Châu Phi, điểm số của 2 Châu Lục này là thấp.



Biểu đồ 5.4. World Map - phân cụm các quốc gia

5.2. Hồi quy tuyến tính đa biến

5.2.1. Mục Đích

Trong phân tích này, chúng ta sẽ thực hiện một mô hình hồi quy tuyến tính đa biến để đánh giá mối quan hệ giữa một biến phụ thuộc và nhiều biến độc lập. Mục tiêu là xác định cách mà các biến độc lập ảnh hưởng đến biến phụ thuộc và đồng thời đánh giá mức độ ảnh hưởng của từng biến.

5.2.2. Dữ liệu

Chúng ta sẽ sử dụng bộ dữ liệu chứa thông tin về điểm số (Score) của các quốc gia, cùng với các biến độc lập khác như GDP, Family, Life Expectancy, Freedom, Generosity, Trust, và Dystopia. Mục tiêu là dự đoán điểm số dựa trên các yếu tố này.

5.2.3. Phương Pháp

Sử dụng phương pháp hồi quy tuyến tính đa biến, chúng ta sẽ xây dựng một mô hình dự đoán điểm số dựa trên tất cả các biến độc lập. Mô hình này sẽ giúp chúng ta hiểu rõ tác động của mỗi yếu tố đối với điểm số và cách chúng tương tác với nhau.

5.2.4. Đầu ra Dự Kiến

Kết quả của mô hình hồi quy sẽ cung cấp hệ số hồi quy cho từng biến độc lập, đồng thời cho phép chúng ta đánh giá độ quan trọng của mỗi yếu tố trong việc dự đoán mức độ hạnh phúc của một quốc gia.

5.2.5. Lợi ích

Hiểu rõ tác động của các yếu tố đối với mức độ hạnh phúc có thể hỗ trợ đưa ra quyết định chính sách và chiến lược phát triển. Mô hình hồi quy đa biến cung cấp cái nhìn chi tiết về cách các yếu tố này đóng góp vào sự khác biệt trong mức sống và hạnh phúc giữa các quốc gia.

5.2.6. Hồi quy tuyến tính

5.2.6.1. Lựa chọn các biến

Quá trình lựa chọn các đặc trưng (feature selection) là một bước quan trọng trong xây dựng mô hình hồi quy. Việc giảm số lượng đặc trưng giúp tăng tốc quá trình huấn luyện mô hình và giảm nguy cơ overfitting. Ta sẽ dùng ma trận tương quan để loại các biến tương quan cao. Tính toán ma trận tương quan giữa các đặc trưng. Loại bỏ các đặc trưng có mức độ tương quan cao (ví dụ, |correlation| > một ngưỡng được đặt trước).

Biểu đồ 5.5. Heatmap - kiểm tra các biến tương quan

Dựa vào biểu đồ tương quan, những màu sắc càng sâu đại diện cho độ tương quan cao. Trong trường hợp này, quan sát được rằng biến Whisker.high và Whisker.low có độ tương đồng lớn với biến Score, với hệ số tương quan đều bằng 1. Điều này chỉ ra rằng cả hai biến Whisker.high và Whisker.low đều có ảnh hưởng trực tiếp lên biến Score.

Dựa trên quan sát này, quyết định loại bỏ cả hai biến Whisker.high và Whisker.low là hợp lý để tránh tình trạng đa cộng tuyến và giữ cho mô hình không bị ảnh hưởng quá mức bởi những biến tương tự. Điều này nhằm mục đích cải thiện sự hiệu quả của mô hình hồi quy bằng cách giữ lại các biến quan trọng và loại bỏ những biến tương quan cao, giúp mô hình trở nên dễ diễn giải và ít nhạy cảm hơn đối với nhiễu từ các biến tương tự.

5.2.6.2. Tách ra feature và target

Quá trình tách tập dữ liệu thành các thành phần feature và target là một phần quan trọng trong quá trình chuẩn bị dữ liệu cho mô hình học máy.

Đầu tiên, ta tiến hành phân chia dữ liệu thành hai tập hợp chính: tập dữ liệu feature (được ký hiệu là X), chứa các biến độc lập hoặc thuộc tính, và tập dữ liệu target (được ký hiệu là Y), chứa biến phụ thuộc. Việc này giúp chúng ta xác định rõ ràng sự phụ thuộc giữa các biến và mục tiêu dự đoán.

Chia thành hai thành phần này là quan trọng để huấn luyện mô hình trên dữ liệu và đánh giá hiệu suất dự đoán của mô hình trên tập kiểm thử hoặc tập dữ liệu mới. Điều này giúp đảm bảo rằng mô hình được học từ các mối quan hệ giữa feature và target một cách hiệu quả.

```
# Chia dữ liệu thành biến phụ thuộc (Y) và biến độc lập (X)
Y = df['Score']
X = df[['GDP', 'Family', 'Life Expectancy', 'Freedom', 'Generosity', 'Trust','Dystopia']]

X = sm.add_constant(X)  # Thêm hệ số chặn (intercept) vào mô hình
```

5.2.6.3 Huấn luyện mô hình

Sau khi phân chia dữ liệu thành tập dữ liệu huấn luyện (X và Y), ta tiến hành xây dựng mô hình hồi quy tuyến tính bằng cách sử dụng thư viện statsmodels. Đầu tiên, ta thêm hệ số chặn (intercept) vào tập dữ liệu độc lập (X) để tạo thành ma trận thiết kế hoàn chỉnh. Sau đó, ta sử dụng phương pháp Ordinary Least Squares (OLS) để ước lượng các tham số của mô hình. Quá trình huấn luyện mô hình OLS này nhằm tối ưu hóa hàm mất mát, trong đó hàm mất mát được định nghĩa bởi sự chênh lệch giữa giá trị dự đoán của mô hình và giá trị thực tế của biến phụ thuộc (Score). Mô hình sẽ điều chỉnh các tham số (bao gồm hệ số chặn và các hệ số của các biến độc lập) để đảm bảo rằng sự chênh lệch này là nhỏ nhất có thể.

Để đánh giá chất lượng của mô hình, ta có thể sử dụng phương thức ***summary()*** của đối tượng mô hình (***mode***) để xem thông tin chi tiết về hệ số hồi quy, giá trị p, R-squared, và các thông số khác.

Phần kết quả của mô hình hồi quy tuyến tính được hiển thị thông qua phương thức **summary()** cho mô hình đã được huấn luyện. Dưới đây là giải thích chi tiết từ kết quả trên:

1. R-squared (R^2): Đánh giá khả năng giải thích của mô hình đối với biến phụ thuộc. Trong trường hợp này, R-squared là 0.995, tức là mô hình giải thích 99.5% sự biến động của biến phụ thuộc.
2. Adjusted R-squared (Adj. R^2): Là một phiên bản điều chỉnh của R-squared để đối phó với số lượng biến độc lập. Trong trường hợp này, nó cũng là 0.995.
3. F-statistic (F-statistic): Kiểm định xem có sự ảnh hưởng chung nào đó từ tất cả các biến độc lập đến biến phụ thuộc hay không. Ở đây, giá trị F-statistic là 4153.
4. Prob (F-statistic): Xác suất của F-statistic, thường được sử dụng để kiểm tra giả thuyết rằng tất cả các hệ số hồi quy đều bằng không. Trong trường hợp này, giá trị rất gần 0, nghĩa là có ít khả năng giả thuyết trên được chấp nhận.
5. Hệ số hồi quy (Coefficients): Đây là các hệ số tương ứng với từng biến độc lập trong mô hình:
    - const (hệ số chặn): 0.0021
    - GDP: 0.9137
    - Family: 1.0247
    - Life Expectancy: 1.1265
    - Freedom: 1.1103
    - Generosity: 0.9311
    - Trust: 1.0648
    - Dystopia: 0.9700
6. P-value (P>|t|): Xác suất của kiểm định t thống kê đối với từng hệ số. Nếu p-value < 0.05, chúng ta có thể bác bỏ giả thuyết rằng hệ số đó bằng 0.
7. Standard Errors: Đo lường độ biến động của hệ số ước lượng.
8. Omnibus, Durbin-Watson, Jarque-Bera, Skew, Kurtosis: Các thống kê khác liên quan đến giả định và tính chất của phân phối của các sai số dự đoán.
9. Prob(Omnibus): Xác suất của kiểm định Omnibus, kiểm định về sự phân phối của các sai số dự đoán.
10. Cond. No. (Condition Number): Đo lường độ nhạy cảm của mô hình đối với biến độc lập, có thể chỉ ra vấn đề về đa cộng tuyến.

Kết quả này cung cấp một bức tranh tổng quan về hiệu suất của mô hình hồi quy tuyến tính và các thông tin đánh giá chất lượng mô hình.

Đây là phương trình hồi quy tuyến tính bội, cho biết cách mỗi biến độc lập đóng góp vào dự đoán giá trị của biến phụ thuộc (Score).

$$Score = 0.0021 + 0.9137 * GDP + 1.0247 * Family + 1.1265 * Life\ Expectancy + 0.9311 * Generosity + 1.0648 * Trust + 0.97 * Dystopia$$

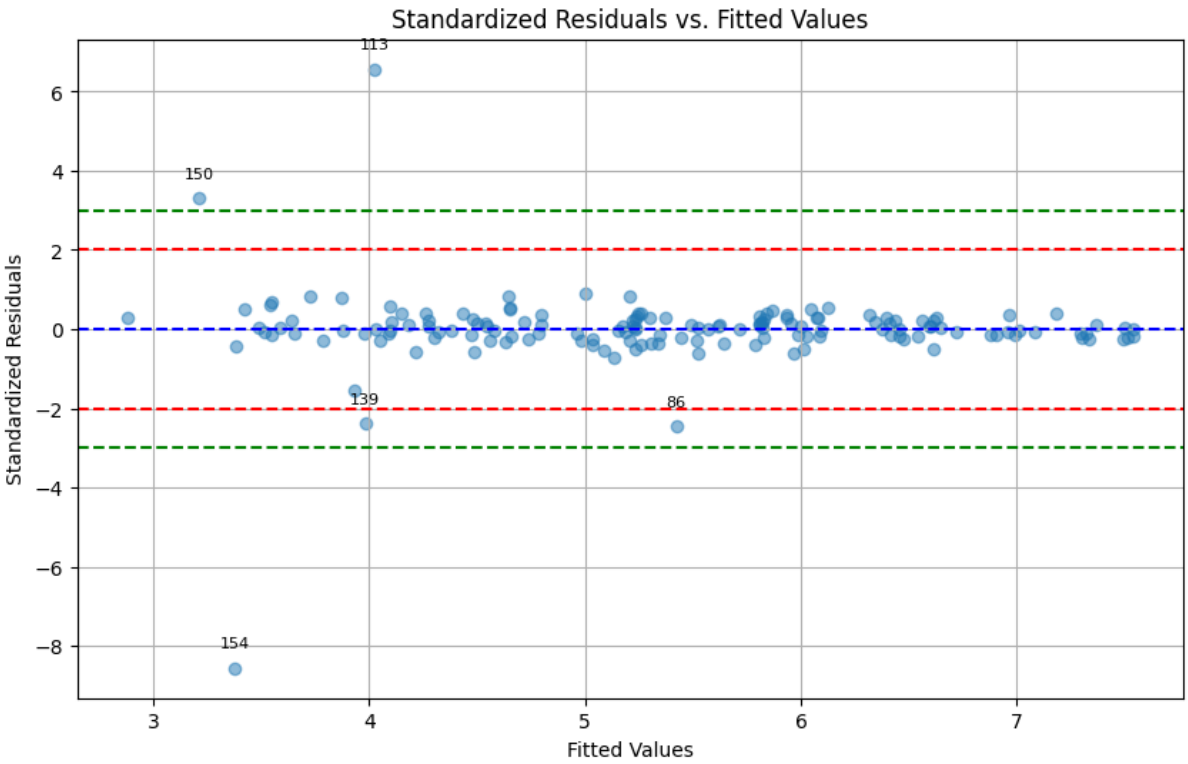
5.2.6.4. Phân tích Đồ thị Residuals chuẩn hóa đối với Mô hình Hồi quy Tuyến tính

Đồ thị residuals chuẩn hóa cung cấp thông tin về sự phân phối của residuals và xác định xem mô hình có đáp ứng đúng các giả định của mô hình hồi quy không.

Mục tiêu phân tích

Xác định sự phân phối và tính chất của residuals chuẩn hóa.

Kiểm tra sự tuân thủ của residuals đối với giả định về phân phối chuẩn và không có xu hướng (homoscedasticity).



Biểu đồ 5.6. Đồ thị Residuals chuẩn hóa

Kết quả và Phân tích

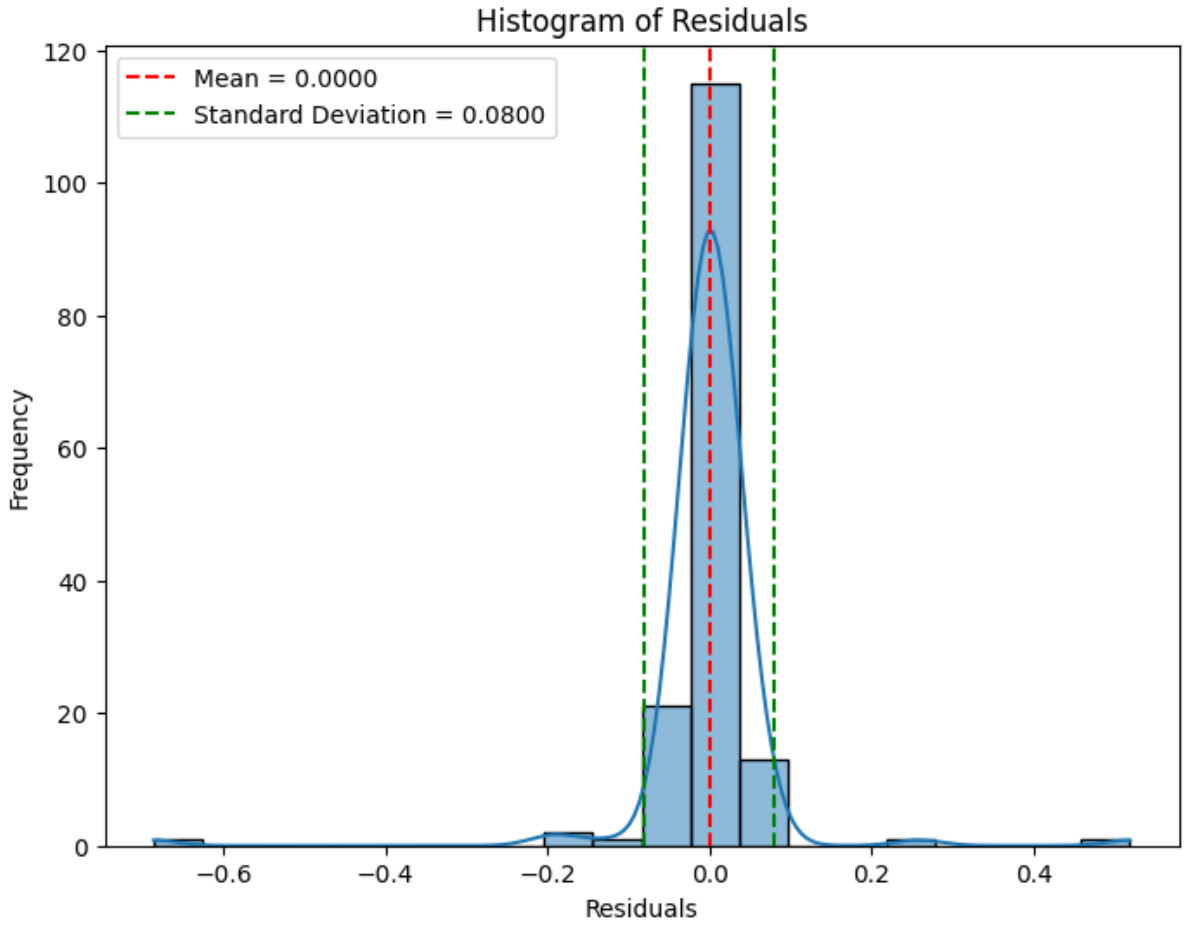
- Đồ thị trên hiển thị sự phân phối của residuals chuẩn hóa dọc theo các giá trị dự đoán (fitted values).

- Các đường giới hạn được thiết lập tại ±2 độ lệch chuẩn và ±3 độ lệch chuẩn để kiểm tra sự phân phối của residuals.
- Các điểm residuals nằm ngoài các giới hạn ±2 và ±3 được đánh số chỉ mục và chú thích trên đồ thị.

Kết luận và Đề xuất

- Đồ thị residuals chuẩn hóa cho thấy rằng phần lớn residuals nằm trong khoảng ±2 độ lệch chuẩn, ngụ ý về sự tuân thủ đối với giả định về phân phối chuẩn.-
- Tuy nhiên, một số điểm residuals vượt quá giới hạn ±2 và ±3, đặt ra câu hỏi về sự không đồng nhất (homoscedasticity) của residuals.
- Dựa trên kết quả này, chúng ta có thể cân nhắc kiểm tra lại mô hình hồi quy tuyến tính để đảm bảo rằng mọi giả định đều được đáp ứng và đề xuất các biện pháp cần thiết để cải thiện chất lượng của mô hình.

5.2.6.5. Kiểm tra giả định phân phối chuẩn của phần dư



Biểu đồ 5.7. Biểu đồ tần số phần dư chuẩn hóa Histogram

Kết quả và Phân tích:

- Đồ thị Histogram của phần dư được hiển thị ở trên.
- Các đường kết hợp với Histogram đại diện cho giá trị trung bình và độ lệch chuẩn của phần dư.
- Phần lớn phần dư tập trung xung quanh giá trị trung bình, và đồ thị có hình dạng gần giống với phân phối chuẩn.
- Các giá trị trung bình và độ lệch chuẩn được hiển thị trên đồ thị để thấy rõ.

Đánh giá và Đề xuất:

- Giá trị trung bình rất gần với 0, ngụ ý rằng phần dư có xu hướng tập trung xung quanh giá trị trung bình.
- Độ lệch chuẩn khá nhỏ, chỉ là 0.0799, ngụ ý rằng phần dư có biên độ nhỏ và không biến động lớn.
- Cả hai giá trị trung bình và độ lệch chuẩn đều là những chỉ số tích cực khi kiểm tra giả định về phân phối chuẩn.
- Dựa trên đồ thị, có vẻ như phần dư có phân phối gần với phân phối chuẩn.
- Đường đỏ là giá trị trung bình của phần dư, và các đường xanh là mức độ độ lệch chuẩn.

Tuy nhiên, việc kiểm tra chi tiết hơn có thể thực hiện các kiểm định thống kê như Kolmogorov-Smirnov hoặc Shapiro-Wilk để xác nhận mức độ tuân thủ với phân phối chuẩn.

Bằng cách này, chúng ta có thể cung cấp đánh giá chi tiết hơn về giả định phân phối chuẩn của phần dư và xác định xem mô hình hồi quy tuyến tính có đáp ứng đúng với giả định này hay không.

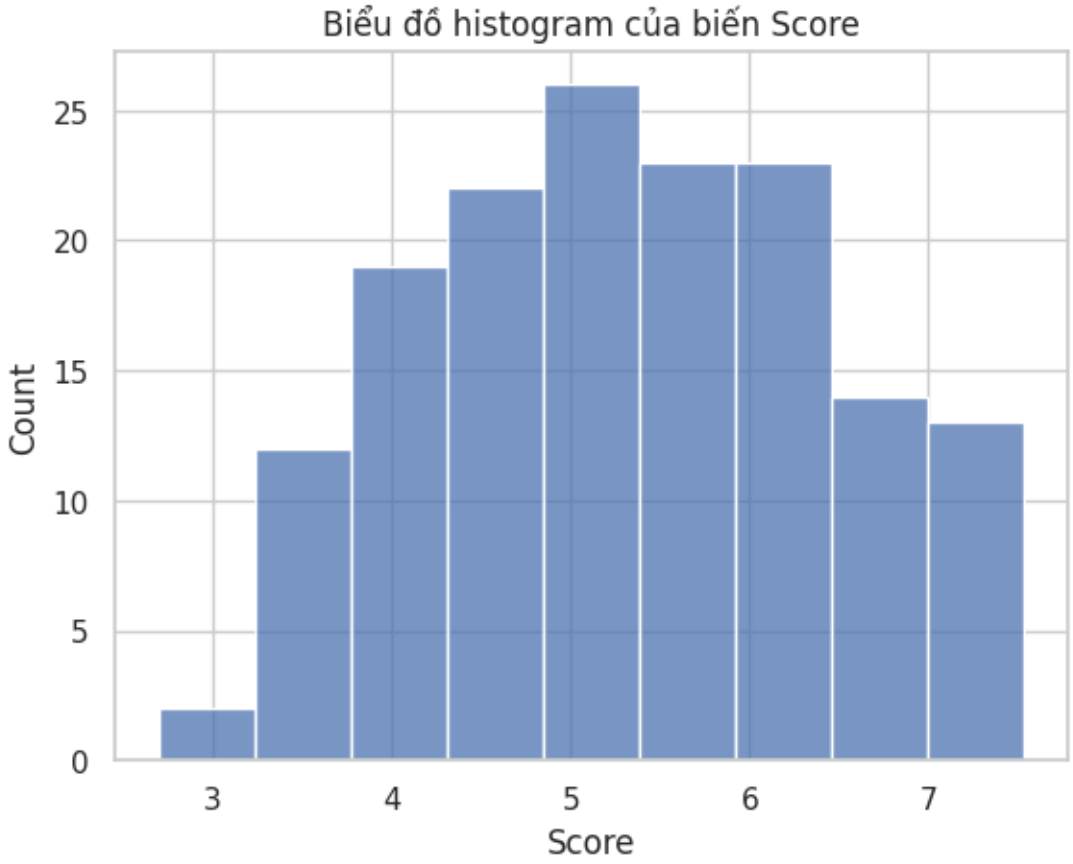
CHƯƠNG VI: KIỂM ĐỊNH GIẢ THUYẾT

6.1. Kiểm định t (t-test)

Định nghĩa: Kiểm định t (t-test) là một công cụ thống kê được sử dụng để so sánh trung bình của hai nhóm dữ liệu khác nhau và xem xét xem liệu sự khác biệt giữa chúng có ý nghĩa thống kê hay không.

Khi thực hiện kiểm định t, chúng ta tính toán một giá trị t dựa trên sự khác biệt giữa trung bình của hai nhóm và sự biến động (độ lệch chuẩn) trong từng nhóm cũng như kích thước của mỗi nhóm. Sau đó, chúng ta so sánh giá trị t này với một ngưỡng xác định (thường là 0.05) để xác định xem sự khác biệt có ý nghĩa thống kê hay không.

Nếu giá trị t vượt qua ngưỡng xác định, chúng ta có đủ bằng chứng để kết luận rằng sự khác biệt giữa hai nhóm là có ý nghĩa thống kê. Nói cách khác, có sự chênh lệch có ý nghĩa giữa trung bình của hai nhóm mà không chỉ là do sự biến động ngẫu nhiên



Biểu đồ 6.1. Histogram của biến Score

.

Trong biểu đồ trên, ta có thể thấy điểm Score của các quốc gia tập trung xung quanh giá trị 5. Ta muốn tìm hiểu xem liệu giá trị trung bình của biến Score có lớn hơn 5 hay không.

Mục tiêu: Kiểm định giá trị trung bình của biến `Score` là lớn hơn hoặc bằng 5

Giả sử chúng ta muốn kiểm định giả thuyết rằng giá trị trung bình của `Score` lớn hơn hoặc bằng 5, với độ tin cậy là 99%; nghĩa là, chúng ta sẽ bác bỏ giả thuyết H0 và ủng hộ giả thuyết thay thế H1 nếu p-value < 0,01.

H0: μScore <= 5

H1: μScore > 5

Kết quả: Vì p-value = 0.0001 < alpha = 0.01 nên ta bác bỏ H0. Từ đây ta có thể kết luận được là giá trị trung bình của biến Score lớn hơn 5

Giải thích: Kết quả của kiểm định t-test ở đây cho thấy p-value = 0.0001 thấp hơn ngưỡng alpha = 0.01 đã chọn.

Khi p-value nhỏ hơn ngưỡng alpha, chúng ta có đủ bằng chứng để bác bỏ giả thuyết không chứng minh (null hypothesis - H0), trong trường hợp này là "giá trị trung bình của biến Score nhỏ hơn hoặc bằng 5". Điều này ngụ ý rằng giá trị trung bình của biến Score thực sự lớn hơn 5 với mức ý nghĩa thống kê.

Do đó, dựa trên kết quả kiểm định, chúng ta có đủ bằng chứng để ủng hộ giả thuyết thay thế (H1) là "giá trị trung bình của biến Score lớn hơn 5" với độ tin cậy 99%.

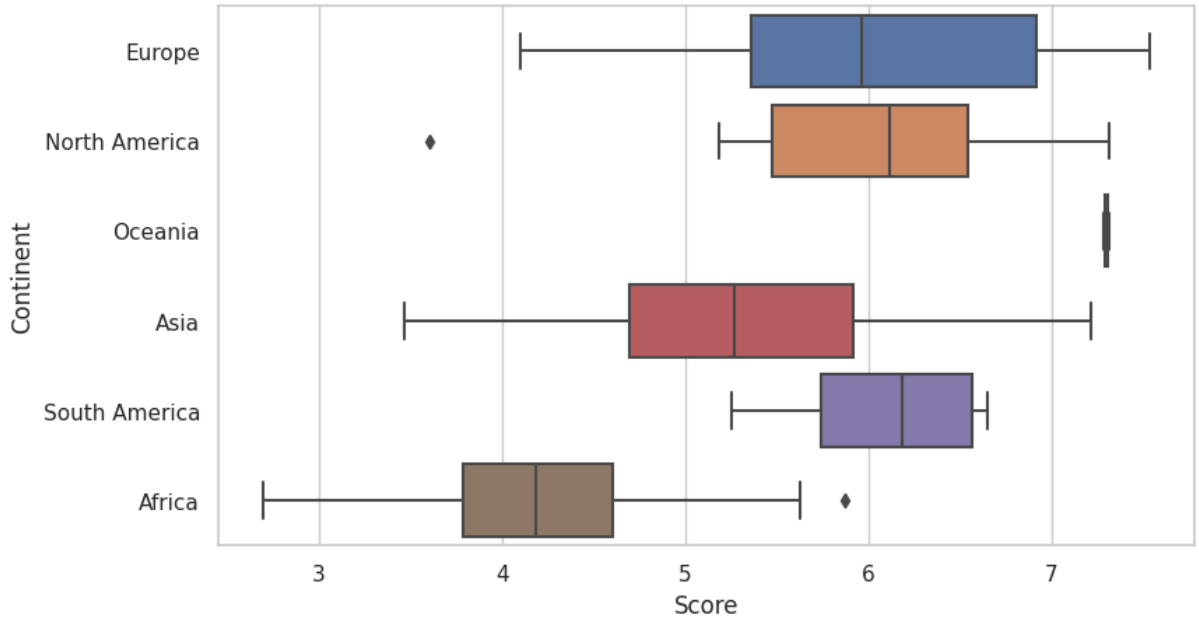
6.2. Kiểm định ANOVA

Định nghĩa: Kiểm định ANOVA (Analysis of Variance) là một phương pháp thống kê dùng để kiểm tra xem có sự khác biệt đáng kể nào đó giữa ba nhóm trở lên hay không. Nó đánh giá sự biến đổi giữa các nhóm và xem xét xem sự khác biệt đó có vượt quá sự biến động ngẫu nhiên mà chúng ta thường mong đợi hay không.

Khi thực hiện kiểm định ANOVA, chúng ta so sánh sự biến động giữa các nhóm với sự biến động trong từng nhóm. Nếu sự biến đổi giữa các nhóm lớn hơn so với sự biến động ngẫu nhiên bên trong từng nhóm, chúng ta có thể kết luận rằng có sự khác biệt ý nghĩa giữa các nhóm.

Ví dụ, giả sử ta muốn biết liệu có sự khác biệt về hiệu suất làm việc giữa ba phương pháp đào tạo A, B và C. Kiểm định ANOVA sẽ giúp xác định xem liệu có sự khác biệt đáng kể nào đó về hiệu suất làm việc giữa ba phương pháp đào tạo này hay không.

Kết quả của kiểm định ANOVA sẽ cung cấp thông tin về sự khác biệt ý nghĩa giữa các nhóm. Nếu kết quả đạt được là ý nghĩa thống kê, chúng ta có thể kết luận rằng ít nhất một trong các phương pháp đào tạo có sự ảnh hưởng đáng kể đến hiệu suất làm việc.



Biểu đồ 6.2. Boxplot giá trị trung bình Score của các châu lục

Trong biểu đồ trên, ta có thể thấy được giữa các châu lục có thể có sự khác biệt đáng kể về giá trị trung bình của biến Score. Ta sẽ thực hiện kiểm định ANOVA để kiểm định giả thuyết này.

Mục tiêu: Kiểm định giả thuyết rằng có sự khác biệt đáng kể về mặt giá trị trung bình giữa các châu lục

H0: Không có sự khác biệt đáng kể về giá trị trung bình của biến Score giữa các châu lục.

H1: Có sự khác biệt đáng kể về giá trị trung bình của biến Score giữa ít nhất một cặp châu lục.

Kết quả:

	sum_sq	df	F	PR(>F)
Continent	98.250351	6.0	24.524326	4.770391e-20
Residual	98.820602	148.0	NaN	NaN

Giải thích:

Kết quả kiểm định ANOVA cung cấp thông tin về sự khác biệt giữa các nhóm (châu lục trong trường hợp này) dựa trên biến đo lường (Score).

Sum of Squares (sum_sq): Đây là tổng bình phương của các sai số hoặc sự khác biệt giữa các giá trị quan sát và giá trị trung bình của mỗi nhóm. Đối với biến 'Continent', tổng bình phương này là 98.25.

Degree of Freedom (df): Độ tự do, là số lượng nhóm hoặc điều kiện (trong trường hợp này là số lượng châu lục), là 6.

F-value (F): Giá trị F, hay còn gọi là chỉ số F, là tỉ lệ giữa sự khác biệt giữa các nhóm và sự biến động bên trong từng nhóm. Trong trường hợp này, giá trị F = 24.52.

PR(>F): Giá trị p, hoặc p-value, là giá trị xác suất. Nó cho biết khả năng rằng sự khác biệt giữa các nhóm là ngẫu nhiên. Trong trường hợp này, p-value rất nhỏ, gần bằng 0 (4.77e-20), thấp hơn rất nhiều so với mức ý nghĩa thống kê thông thường (0.05).

Kết quả này cho thấy rằng giữa các châu lục có sự khác biệt đáng kể về giá trị trung bình của biến Score. Với giá trị p-value rất thấp, chúng ta có đủ bằng chứng để bác bỏ giả thuyết rằng không có sự khác biệt đáng kể giữa các châu lục. Điều này ngụ ý rằng ít nhất một cặp châu lục có sự khác biệt đáng kể về giá trị trung bình của biến Score.

6.3. Kiểm định Tukey's HSD (Honestly Significant Difference)

Định nghĩa: Kiểm định Tukey's HSD (Honestly Significant Difference) là một phương pháp thống kê dùng để so sánh trung bình của các nhóm một cách chi tiết và cụ thể sau khi kiểm định ANOVA (Analysis of Variance) đã cho thấy có sự khác biệt chung giữa các nhóm.

Khi kiểm định ANOVA chỉ ra rằng có sự khác biệt chung giữa các nhóm, Tukey's HSD sẽ giúp xác định chính xác các cặp nhóm nào có trung bình khác biệt đáng kể với nhau.

Ví dụ, giả sử bạn đã thực hiện một thí nghiệm với ba hoặc nhiều hơn nhóm, và kết quả từ kiểm định ANOVA cho thấy có sự khác biệt ý nghĩa giữa các nhóm. Khi đó, bạn có thể sử dụng kiểm định Tukey's HSD để xác định chính xác nhóm nào khác biệt nhau.

Kết quả từ Tukey's HSD sẽ cung cấp thông tin về các cặp nhóm có sự khác biệt đáng kể về trung bình. Nếu giữa các cặp nhóm, Tukey's HSD chỉ ra có sự khác biệt ý nghĩa, điều này sẽ giúp bạn hiểu rõ hơn về mức độ khác biệt cụ thể giữa các nhóm sau khi đã thực hiện kiểm định ANOVA.

Mục tiêu: Thực hiện kiểm định Tukey's HSD sẽ giúp xác định chính xác các cặp châu lục nào có trung bình biến Score khác biệt đáng kể với nhau.

Kết quả:

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Africa	Asia	1.0644	0.0	0.5465	1.5822	True
Africa	Europe	1.899	0.0	1.3618	2.4362	True
Africa	North America	1.7887	0.0	1.0392	2.5382	True
Africa	Oceania	3.0595	0.0	1.2935	4.8255	True
Africa	South America	1.8591	0.0	1.0034	2.7148	True
Africa	other	1.0395	0.8697	-1.4308	3.5098	False
Asia	Europe	0.8347	0.0001	0.3003	1.3691	True
Asia	North America	0.7243	0.0642	-0.0232	1.4719	False
Asia	Oceania	1.9951	0.0159	0.2299	3.7603	True
Asia	South America	0.7947	0.0862	-0.0592	1.6487	False
Asia	other	-0.0249	1.0	-2.4945	2.4448	False
Europe	North America	-0.1103	0.9995	-0.8714	0.6507	False
Europe	Oceania	1.1605	0.4453	-0.6105	2.9314	False
Europe	South America	-0.0399	1.0	-0.9058	0.8259	False
Europe	other	-0.8595	0.9441	-3.3333	1.6142	False
North America	Oceania	1.2708	0.3837	-0.5757	3.1173	False
North America	South America	0.0704	1.0	-0.941	1.0817	False
North America	other	-0.7492	0.9742	-3.2776	1.7792	False
Oceania	South America	-1.2004	0.4859	-3.0925	0.6917	False
Oceania	other	-2.02	0.4076	-5.0116	0.9716	False
South America	other	-0.8196	0.9623	-3.3815	1.7423	False

Giải thích:

Kết quả kiểm định Tukey's HSD cung cấp thông tin về sự khác biệt giữa các cặp châu lục về giá trị trung bình của biến Score sau khi đã thực hiện kiểm định ANOVA.

group1 và group2: Các nhóm châu lục được so sánh với nhau.

meandiff: Độ chênh lệch trung bình giữa các nhóm.

p-adj: Giá trị p đã được điều chỉnh (adjusted p-value) để điều chỉnh rủi ro liên quan đến nhiều lần thử nghiệm.

lower và upper: Giới hạn dưới và giới hạn trên của khoảng tin cậy 95% cho sự khác biệt trung bình giữa các nhóm.

reject: True hoặc False, chỉ ra liệu chúng ta có thể bác bỏ giả thuyết không có sự khác biệt giữa các nhóm hay không, dựa trên giá trị p đã điều chỉnh.

Thông qua bảng kết quả, chúng ta có thể kết luận rằng:

- Các cặp châu lục như Africa - Asia, Africa - Europe, Africa - North America, Africa - Oceania, Africa - South America, Asia - Europe và Asia - Oceania đều có sự khác biệt đáng kể về giá trị trung bình của biến Score.
- Trong khi đó, các cặp nhóm khác như Asia-North America, Asia-South America, Europe-North America, Europe-Oceania, Europe-South America, North America-Oceania, North America-South America, Oceania-South America, và South America - other không có sự khác biệt đáng kể về giá trị trung bình của biến Score.

6.4. Kiểm định Shapiro-Wilk

Định nghĩa: Kiểm định Shapiro-Wilk là một phương pháp thống kê được sử dụng để kiểm tra xem một mẫu dữ liệu có tuân theo phân phối chuẩn (phân phối Gaussian) hay không. Ý tưởng cơ bản của kiểm định này là so sánh giữa đặc tính phân phối của mẫu dữ liệu với phân phối chuẩn được kỳ vọng.

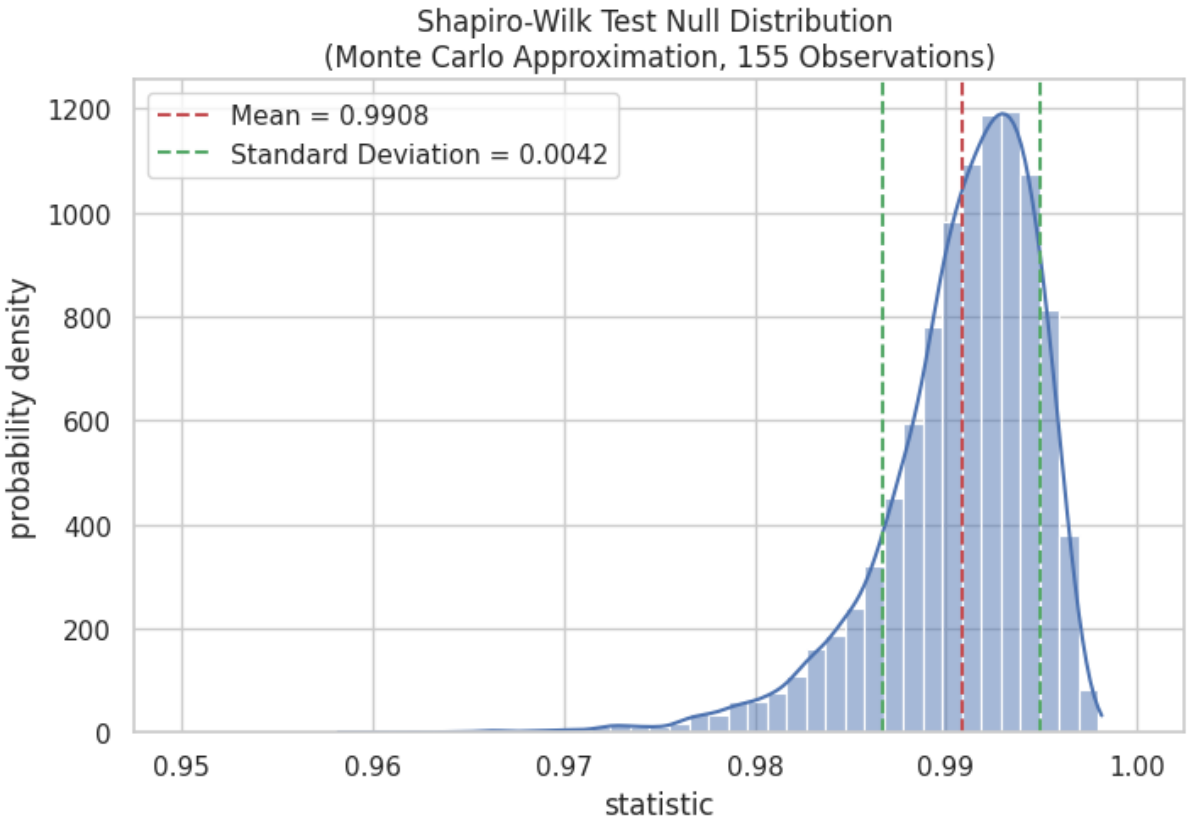
Quá trình kiểm định này hoạt động bằng cách so sánh các giá trị được quan sát trong mẫu dữ liệu với các giá trị dự đoán từ một phân phối chuẩn lý tưởng. Nếu mẫu dữ liệu không tuân theo phân phối chuẩn, kiểm định Shapiro-Wilk sẽ cung cấp cho chúng ta một độ đo (statistic) và một giá trị p (p-value) để xác định mức độ mà dữ liệu không tuân theo phân phối chuẩn.

Khi giá trị p nhỏ hơn một ngưỡng xác định trước (thường là 0.05), chúng ta có đủ bằng chứng để bác bỏ giả thuyết rằng mẫu dữ liệu tuân theo phân phối chuẩn. Điều này có nghĩa là chúng ta có thể kết luận rằng mẫu dữ liệu có một đặc điểm phân phối khác biệt so với phân phối chuẩn.

Mục tiêu: Với độ tin cậy alpha = 0.05, ta cần kiểm định giả thuyết:

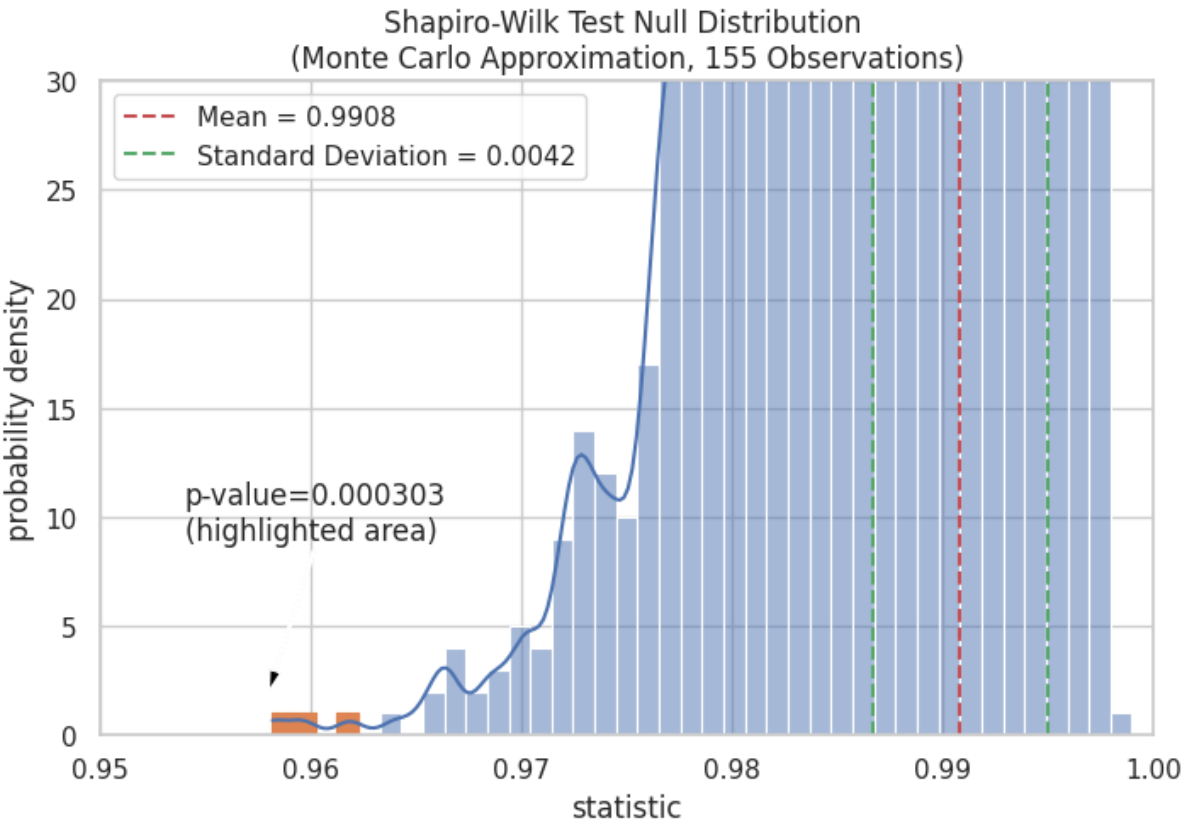
H0: Biến Freedom có dạng phân phối chuẩn

H1: Biến Freedom không có dạng phân phối chuẩn



Biểu đồ 6.3. Histogram kiểm định phân phối chuẩn

Việc kiểm định được thực hiện bằng cách so sánh giá trị thống kê với phân phối được hình thành theo giả thuyết không với các trọng số lấy từ phân phối chuẩn. Đối với kiểm định phân phối chuẩn này, phân phối của giả thuyết không (null hypothesis) không dễ tính toán chính xác, do đó nó thường được xấp xỉ bằng phương pháp Monte Carlo, nghĩa là lấy nhiều mẫu có cùng kích thước với x từ phân phối chuẩn và tính toán các giá trị của thống kê cho mỗi mẫu.



Biểu đồ 6.4. Histogram kiểm định p-value

Nếu p-value "nhỏ" - nghĩa là, việc lấy mẫu dữ liệu từ một quần thể có phân bố chuẩn tạo ra giá trị thống kê lớn như vậy có xác suất vô cùng nhỏ - điều này có thể được coi là bằng chứng chống lại H0 và ủng hộ giả thuyết thay thế H1 : các trọng số không được rút ra từ phân phối chuẩn.

Kết quả: Kết quả của kiểm định Shapiro-Wilk cho biết giá trị p-value rất nhỏ, cụ thể là p-value = 0.000303.

Giải thích: Khi p-value nhỏ hơn ngưỡng ý nghĩa thống kê trước đặt ra (ở đây là 0.05), chúng ta có đủ bằng chứng để bác bỏ giả thuyết H0, tức là giả định rằng biến Freedom tuân theo phân phối chuẩn.

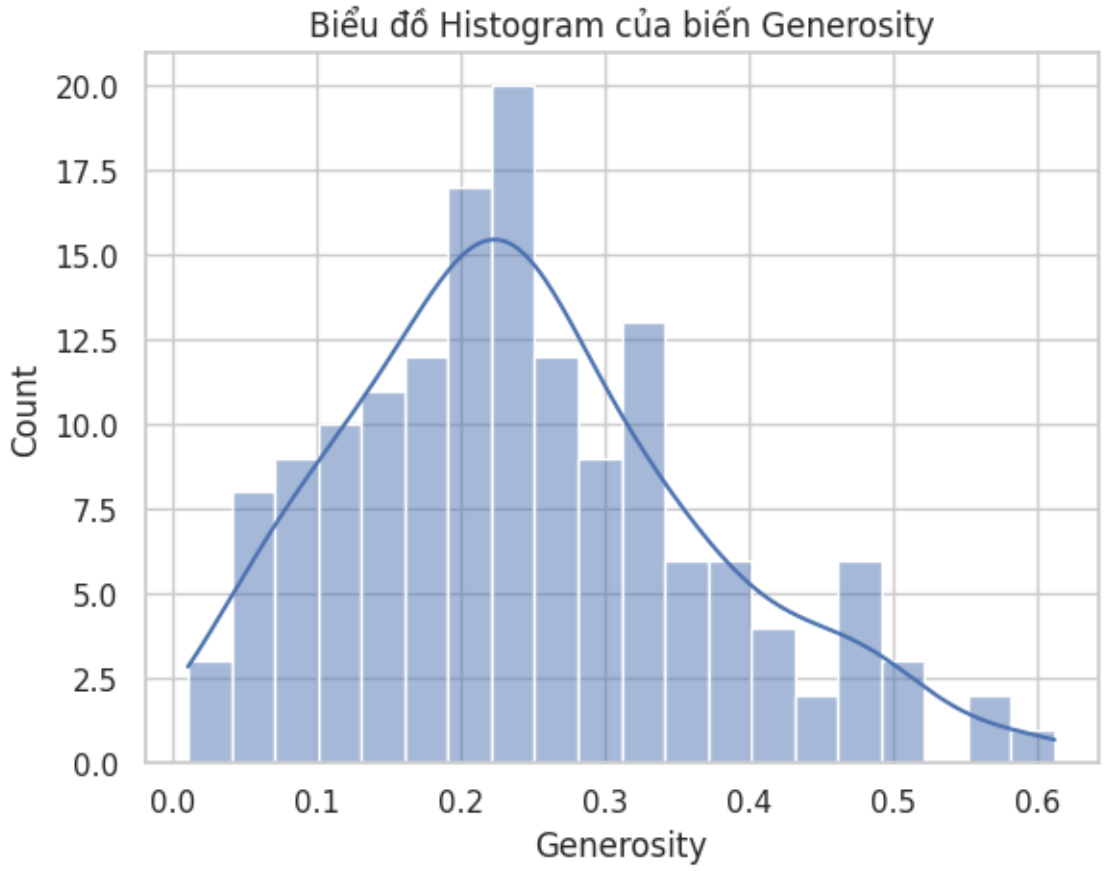
Do đó, dựa trên kết quả kiểm định, chúng ta có đủ thông tin để kết luận rằng có bằng chứng đủ mạnh để phủ nhận giả thuyết rằng biến Freedom có dạng phân phối chuẩn. Thay vào đó, có dấu hiệu cho thấy rằng dữ liệu không tuân theo phân phối chuẩn theo cách chúng ta hiểu về phân phối Gaussian hoặc phân phối chuẩn.

6.5. Kiểm định Skewness

Định nghĩa: Kiểm định skewness là một phương pháp thống kê được sử dụng để kiểm tra xem một phân phối dữ liệu có đối xứng hay không. Skewness (độ lệch) đo lường mức độ mà dữ liệu phân phối không đối xứng so với phân phối chuẩn hoặc đối xứng.

Một phân phối đối xứng có skewness gần bằng 0, trong khi skewness dương cho thấy đỉnh của phân phối nằm ở bên trái và đuôi dài về phía bên phải, và skewness âm cho thấy đỉnh của phân phối nằm ở bên phải và đuôi dài về phía bên trái.

Kiểm định skewness thường sử dụng các phương pháp thống kê như kiểm định z-score hoặc sử dụng các giả định để xác định xem mức độ lệch của dữ liệu có ý nghĩa thống kê hay không.

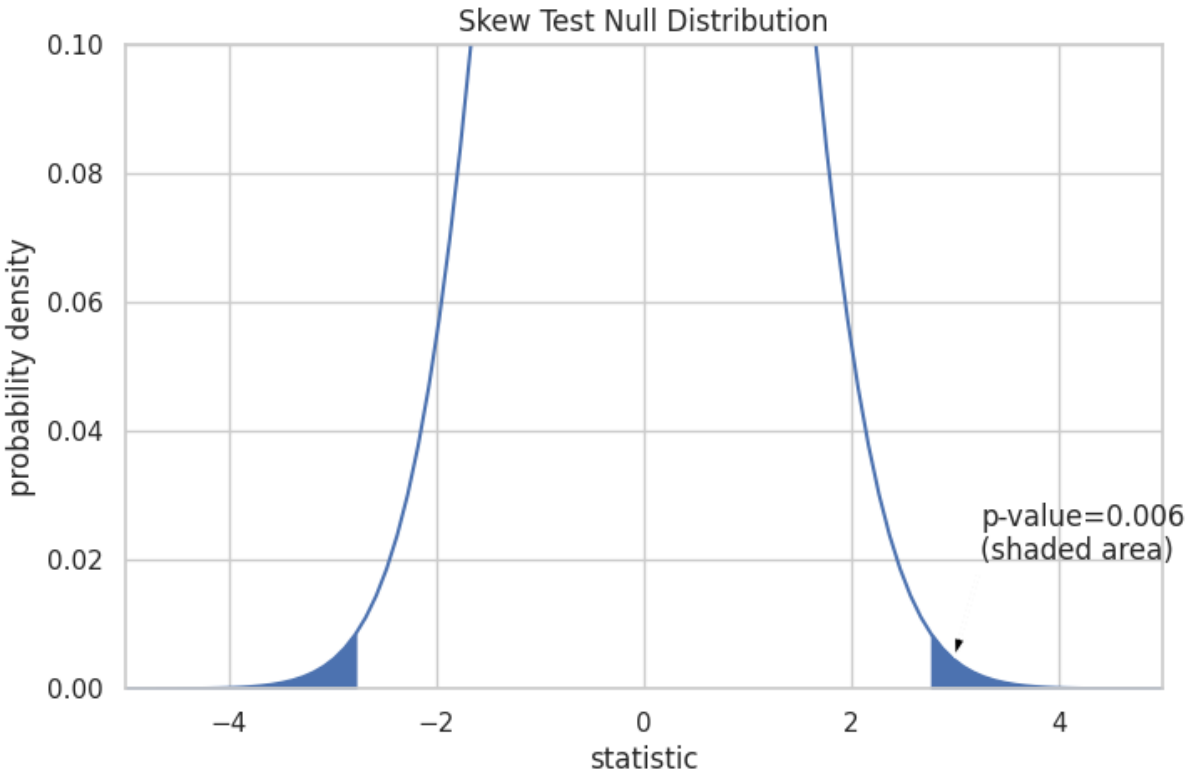


Biểu đồ 6.5. Histogram của biến Generosity

Mục tiêu: Mục tiêu: Dựa trên biểu đồ, chúng ta sử dụng kiểm định độ lệch (skewness) để đánh giá tính đối xứng của phân phối biến Generosity.

H0: Phân phối của biến Generosity là đối xứng.

H1: Phân phối của biến Generosity không đối xứng.



Biểu đồ 6.6. Plot phân phối của biến Generosity

Kết quả: p-value = 0.0069 từ kiểm định skewness cho biết giá trị p-value nhỏ hơn mức ý nghĩa alpha = 0.05. Từ đó, ta có thể bác bỏ giả thuyết H0: Phân phối của biến Generosity là đối xứng.

Giải thích: Kết quả kiểm định skewness cho biến Generosity có giá trị p-value = 0.0069. Khi so sánh giá trị p-value này với ngưỡng ý nghĩa thống kê là alpha = 0.05, chúng ta quyết định bác bỏ giả thuyết H0 về tính đối xứng của phân phối dữ liệu. Thay vào đó, dữ liệu cho thấy mức độ lệch, có sự không đối xứng đáng kể so với phân phối chuẩn hoặc đối xứng.

ĐÁNH GIÁ

Họ và tên	Nội dung	Đóng góp
Trương Thị Hồng Mai	Tổng quan đề tài Phân tích theo châu lục	*100%*
Đặng Đại Lợi	Machine Learning Chỉnh sửa-hoàn thiện báo cáo phần Word	100%
Võ Ngọc Mỹ Kim	Tiền xử lý dữ liệu Phân tích theo châu lục Phân tích tổng quan	100%
Nguyễn King	Kiểm định giả thuyết	100%
Doãn Phương Hà My	Tiền xử lý dữ liệu Phân tích theo châu lục Phân tích tổng quan	100%

TÀI LIỆU THAM KHẢO

[1]. Freedom House. (2023). Freedom in the World — Haiti Country Report. Retrieved from https://freedomhouse.org/country/haiti

[2]. Python Software Foundation. (n.d.). Python.org. Retrieved from https://www.python.org/

[3]. Matplotlib. (n.d.). Matplotlib: Python Plotting. Retrieved from https://matplotlib.org/

[4]. Seaborn Development Team. (n.d.). Seaborn: Statistical Data Visualization. Retrieved from https://seaborn.pydata.org/

[5]. Singh, A. (n.d.). World Happiness Report 2021. Kaggle. Retrieved from https://www.kaggle.com/datasets/ajaypalsinghlo/world-happiness-report-2021

[6]. Kaggle. (n.d.). Kaggle: Your Machine Learning and Data Science Community. Retrieved from https://www.kaggle.com/