**MỤC LỤC**

LỜI NÓI ĐẦU…………………………………………………………………….. 3

PHẦN 1: GIỚI THIỆU BÀI TOÁN………………………………………………. 4

PHẦN 2: TIỀN XỬ LÝ DỮ LIỆU……………………………………………….. 5

2.1. Xử lý dữ liệu thiếu……………………………………………………... 6

2.2 Xử lý dữ liệu ngày tháng……………………………………………….. 6

PHẦN 3: TRỰC QUAN HÓA DỮ LIỆU……………………………………… 8

3.1. Biểu đồ Heatmap……………………………………………………… 8

3.2. Biểu đồ Boxplot………………………………………………………. 9

3.3. Biểu đồ Scatter matrix………………………………………………... 10

3.4. Biểu đồ Bar chart……………………………………………………... 11

3.4. Biểu đồ cột xếp chồng………………………………………………... 11

3.6. Biểu đồ Sunburst……………………………………………………... 13

3.7. Biểu đồ Scatter geo…………………………………………………... 15

3.8. Biểu đồ cột…………………………………………………………… 16

KẾT LUẬN……………………………………………………………………… 19

TÀI LIỆU THAM KHẢO………………………………………………………. 20

**LỜI NÓI ĐẦU**

Lời đầu tiên nhóm chúng em xin chân thành gửi lời cảm ơn đến thầy giáo …đã đồng hành, hướng dẫn cùng chúng em để hoàn thành đề tài này.

Trực quan hóa dữ liệu là một phần không thể thiếu trong việc hiểu và phân tích thông tin, giúp chúng ta nhanh chóng nhận ra các mô hình, xu hướng, và ngoại lệ từ các bộ dữ liệu lớn và phức tạp. Trong bối cảnh nền kinh tế toàn cầu hiện nay đang bị ảnh hưởng bởi dịch Covid-19, tỉ lệ thất nghiệp là một trong những chỉ số kinh tế quan trọng, phản ánh trạng thái của thị trường lao động và khả năng sinh kế của người dân. Đặc biệt, ở Ấn Độ, một quốc gia đa dạng với sự khác biệt lớn về văn hóa, kinh tế giữa các bang, việc theo dõi và phân tích tỉ lệ thất nghiệp trở nên cực kỳ quan trọng để đánh giá sức khỏe kinh tế và xã hội của quốc gia.

Đề tài trực quan hóa tỉ lệ thất nghiệp của các bang ở Ấn Độ không chỉ mang lại cái nhìn tổng quan về tình hình thất nghiệp ở các bang khác nhau mà còn giúp chúng ta tìm hiểu sâu hơn về những yếu tố ảnh hưởng đến thị trường lao động và các chính sách cần thiết để thúc đẩy việc làm. Thông qua việc áp dụng các kỹ thuật trực quan hóa dữ liệu tiên tiến, đề tài này sẽ cung cấp một cái nhìn mới mẻ và chính xác về cách tỉ lệ thất nghiệp phân bổ ở Ấn Độ trong thời kỳ Covid-19, từ đó hỗ trợ cho cả người hoạch định chính sách và công chúng trong việc hiểu và đối phó với các thách thức về lao động và việc làm.

Qua đề tài này, chúng tôi mong muốn đóng góp vào quá trình nghiên cứu và phát triển các phương pháp trực quan hóa dữ liệu, đặc biệt là đối với dữ liệu trong lĩnh vực kinh tế nói chung và việc làm nói riêng. Hy vọng rằng công trình của chúng tôi sẽ mở ra những hướng đi mới và tạo động lực cho sự phát triển của các nghiên cứu tiếp theo trong cộng đồng học thuật và thực tiễn áp dụng.

**PHẦN 1: GIỚI THIỆU BÀI TOÁN**

Bản báo cáo này chúng em tổng quan trực quan hóa dữ liệu cho bộ dữ liệu. Mục tiêu của đề tài là để biết được tác động của Covid19 đến thị trường việc làm ở các bang tại Ấn Độ. Bang nào ít bị ảnh hưởng và bang nào bị ảnh hưởng nhiều.

Thông tin dataset: Bộ dữ liệu nhóm chúng em dùng trong dự án này bao gồm thông tin về các bang, ngày tháng được theo dõi, tỷ lệ và số lượng người thất nghiệp, Tỉ lệ người trong độ tuổi lao động.

Dataset gồm có 268 bản ghi và 9 thuộc tính, bao gồm:

* States = Các bang ở Ấn Độ
* Date = date which the unemployment rate observed
* Frequency = Tần suất đo (Hàng tháng)
* Estimated Unemployment Rate (%) = Tỷ lệ người thất nghiệp ở mỗi bang
* Estimated Employed = Số lượng người thất nghiệp
* Estimated Labour Participation Rate (%) = Tỉ lệ người trong độ tuổi lao động từ 16 tuổi đến 64 tuổi
* Region: Khu vực
* Longitude: Kinh độ
* Latitude: Vĩ độ

**PHẦN 2: TIỀN XỬ LÝ DỮ LIỆU**

Trước khi tiền xử lý ta phải khai báo thư viện và tải tập dữ liệu lên bằng câu lệnh.

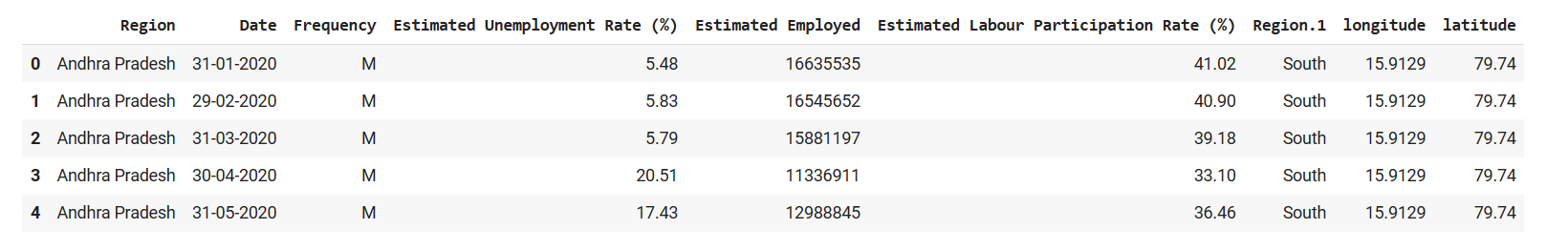
|  |
| --- |
| import numpy as np  import pandas as pd  import matplotlib.pyplot as plt  import seaborn as sns  import calendar  import datetime as dt  import plotly.io as pio  pio.templates  import plotly.express as px  import plotly.graph\_objects as go  import plotly.figure\_factory as ff  from IPython.display import HTML |

|  |
| --- |
| df = pd.read\_csv('Unemployment\_Rate\_upto\_11\_2020.csv') |

Ta kiểm tra xem dữ liệu đã được tải lên chưa bằng câu lệnh

|  |
| --- |
| df.head() |

Kết quả thu được:



**2.1. Xử lý dữ liệu thiếu**

Ta kiểm tra xem liệu bộ dữ liệu có bị thiếu không:

|  |
| --- |
| df.isnull().sum() |

Kết quả thu được:

Region 0

Date 0

Frequency 0

Estimated Unemployment Rate (%) 0

Estimated Employed 0

Estimated Labour Participation Rate (%) 0

Region.1 0

longitude 0

latitude 0

Như vậy bộ dữ liệu không chứa dữ liệu thiếu. Ta có thể bỏ qua bước này.

**2.2 Xử lý dữ liệu ngày tháng**

Chuyển đổi kiểu dữ liệu của cột 'Date' thành datetime, định dạng ngày tháng trong dữ liệu là ngày trước tháng:

|  |
| --- |
| df['Date'] = pd.to\_datetime(df['Date'],dayfirst=True) |

Tạo cột mới 'Month' từ cột 'Date', trích xuất tháng từ cột 'Date' và lưu vào cột mới 'Month':

|  |
| --- |
| df['Month'] = df['Date'].dt.month |

Chuyển đổi cột 'Month' thành số nguyên. Chuyển giá trị tháng được trích xuất thành số nguyên và lưu vào cột mới 'Month\_int':

|  |
| --- |
| df['Month\_int'] = df['Month'].apply(lambda x : int(x)) |

Tạo cột mới 'Month\_name' từ 'Month\_int', chuyển đổi số tháng sang tên tháng viết tắt và lưu vào cột 'Month\_name':

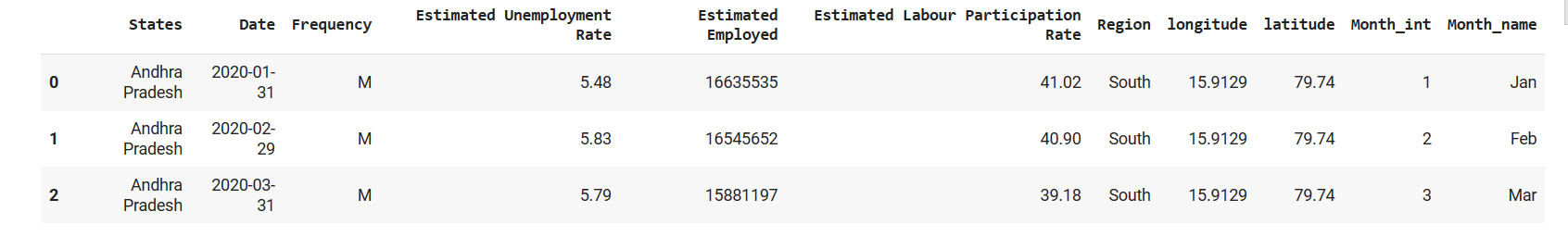
|  |
| --- |
| df['Month\_name'] = df['Month\_int'].apply(lambda x: calendar.month\_abbr[x]) |

Việc sử dụng calendar.month\_abbr[x] giúp chuyển số tháng thành tên tháng tương ứng bằng tiếng Anh.

Ta loại bỏ cột 'Month' khỏi DataFrame, vì thông tin này chúng em đã thay thế bằng 'Month\_int' và 'Month\_name':

|  |
| --- |
| df.drop(columns='Month', inplace=True) |

Ta kiểm tra lại dữ liệu:



**PHẦN 3: TRỰC QUAN HÓA DỮ LIỆU**

Về định nghĩa trực quan hóa dữ liệu là quá trình biểu diễn thông tin số liệu bằng các đồ họa, biểu đồ hoặc hình ảnh để truyền đạt một cách hiệu quả nhất. Mục đích của việc trực quan hóa dữ liệu là làm cho dữ liệu trở nên dễ hiểu và truyền đạt thông tin một cách rõ ràng và trực quan hơn cho người xem.

**3.1. Biểu đồ Heatmap**

1. Vì sao mối tương quan giữa longtitude(vị trí địa lý của khu vực, được đo bằng vĩ độ) với Estimated Labour Participation Rate(Tỉ lệ người trong độ tuổi lao động từ 16 tuổi đến 64 tuổi) lại cao ?

Để trả lời cho câu hỏi này nhóm chúng em sẽ sử dụng biểu đồ heatmap:

|  |
| --- |
| heat\_maps = df[['Estimated Unemployment Rate',  'Estimated Employed', 'Estimated Labour Participation Rate',  'longitude', 'latitude', 'Month\_int']]  heat\_maps = heat\_maps.corr()  plt.figure(figsize=(10,6))  sns.set\_context('notebook',font\_scale=1)  sns.heatmap(heat\_maps, annot=True,cmap='summer'); |

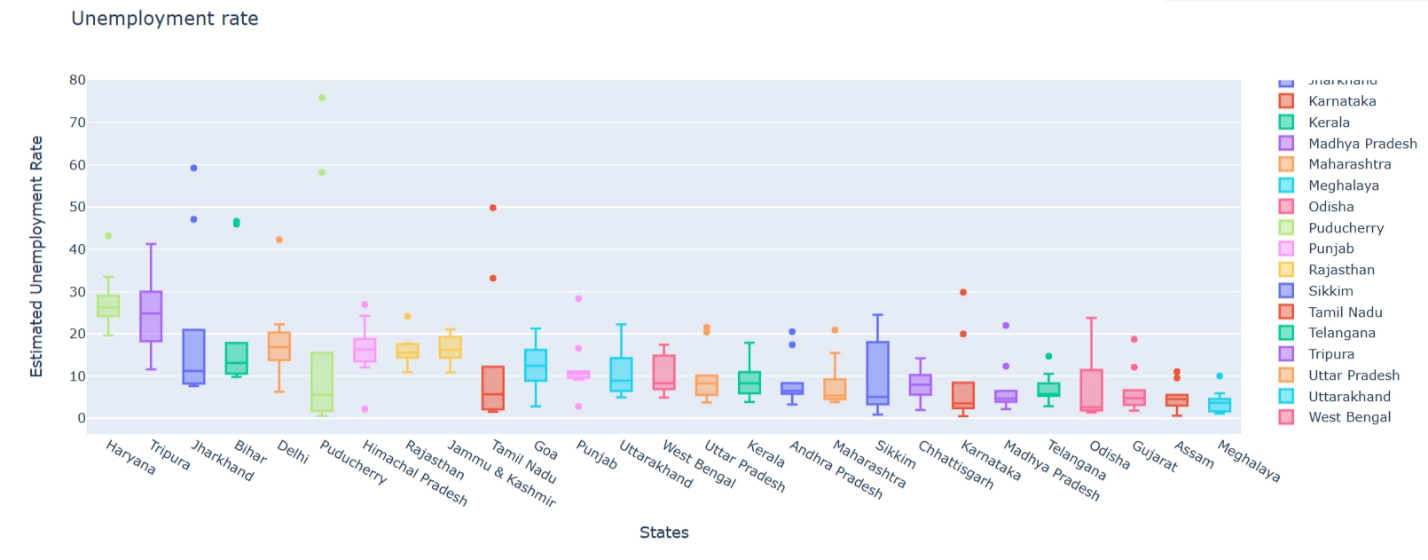


* Vị trí địa lý có thể ảnh hưởng đến mức độ tham gia vào lực lượng lao động trong các khu vực (các tỉnh thành phố lớn trung tâm kinh tế, các thành phố có nhiều khu công nghiệp)

**3.2. Biểu đồ Boxplot**

2. Có những bang nào xuất hiện nhiều outliers (các giá trị ngoại lệ) trong dữ liệu thất nghiệp không? Các outliers này có thể là do những yếu tố nào?

|  |
| --- |
| fig = px.box(df,x='States',y='Estimated Unemployment Rate',color='States',title='Unemployment rate',template='plotly')  fig.update\_layout(xaxis={'categoryorder':'total descending'})  fig.show() |



Dựa vào biểu đồ ta thấy tỉ lệ thất nghiệp của các bang được phân bổ khá đồng đều. Nhưng ta có thể thấy hầu hết các bang đều có một vài điểm ngoại lệ. Các ngoại lệ có thể là do dữ liệu của chúng em là tỉ lệ thất nghiệp có bao gồm dữ liệu từ Covid-19 nên những điểm dữ liệu ấy biểu hiện cho các tháng tỉ lệ cao bất thường vì dịch.

**3.3. Biểu đồ Scatter matrix**

3. Có mối quan hệ nào giữa các biến kinh tế và vùng miền của Ấn Độ?

Với câu hỏi này nhóm sử dụng biểu đồ ma trận phân tán:

|  |
| --- |
| fig = px.scatter\_matrix(df,template='plotly',  dimensions=['Estimated Unemployment Rate','Estimated Employed',  'Estimated Labour Participation Rate'],  color='Region')  fig.show() |



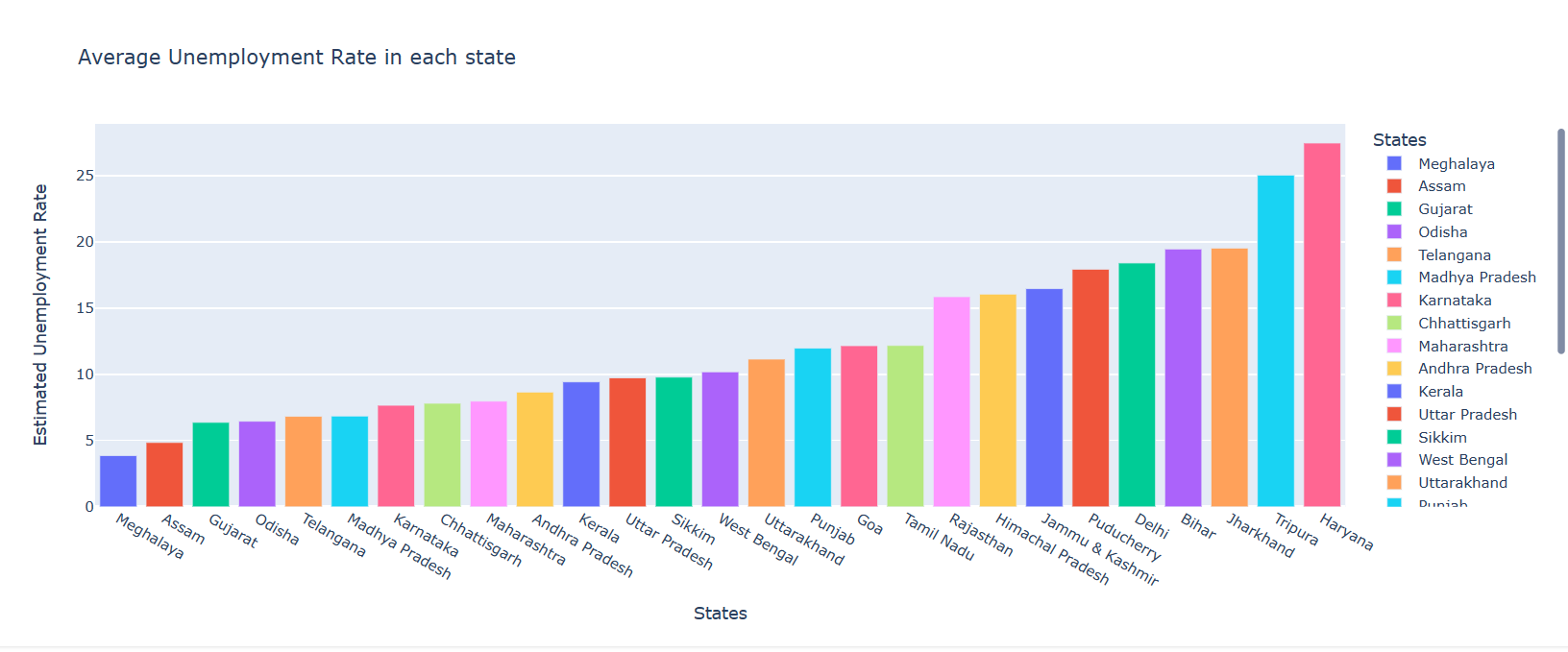
* Có tổng cộng 5 màu tương ứng với 5 vùng, mỗi vùng sẽ có một màu khác nhau
* Nhìn vào biểu đồ ta thấy, vùng đông bắc, tỉ lệ người trong độ tuổi người lao động nằm trong khoảng từ 40 -> 60% và số lượng và số lượng người lao động trong khoảng 2-3 triệu người

=> Ta có thể thấy vùng này có khá ít người so với mặt bằng chung nhưng tỷ lệ người trong độ tuổi lao động lại là lớn nhất

**3.4. Biểu đồ Bar chart**

4. Bang nào có tỷ lệ thất nghiệp trung bình cao nhất dựa trên biểu đồ bar chart?

|  |
| --- |
| plot\_ump = df[['Estimated Unemployment Rate','States']]  df\_unemp = plot\_ump.groupby('States').mean().reset\_index()  df\_unemp = df\_unemp.sort\_values('Estimated Unemployment Rate')  fig = px.bar(df\_unemp, x='States',y='Estimated Unemployment Rate',color='States',  title='Average Unemployment Rate in each state',template='plotly')  fig.show() |



Từ biểu đồ bar chart ta dễ dàng thấy được rằng bang Haryana là bang có tỉ lệ thất nghiệp trung bình cao nhất.

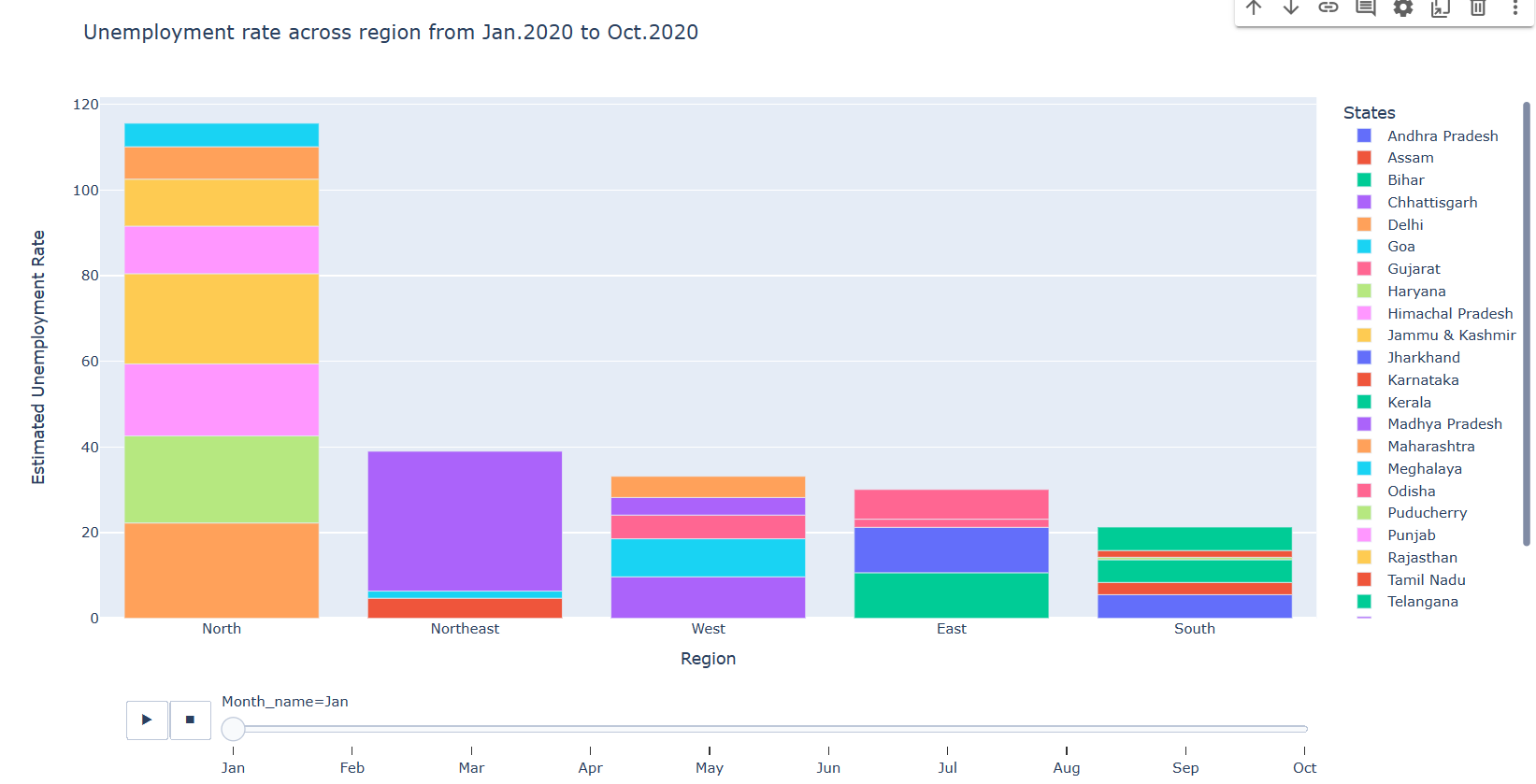
**3.5. Biểu đồ cột xếp chồng**

5.Có sự biến động như thế nào về tỷ lệ thất nghiệp qua các khu vực (Region) từ tháng 1/2020 đến tháng 10/2020?

Vấn đề này nhóm sử dụng biểu đồ cột xếp chồng chạy từ tháng 1 đếns tháng 10:

|  |
| --- |
| fig = px.bar(df, x='Region',y='Estimated Unemployment Rate',animation\_frame = 'Month\_name',color='States',  title='Unemployment rate across region from Jan.2020 to Oct.2020', height=700,template='plotly')  fig.update\_layout(xaxis={'categoryorder':'total descending'})  fig.layout.updatemenus[0].buttons[0].args[1]["frame"]["duration"] = 2000  fig.show() |

Tham số animation\_frame = 'Month\_name' được nhóm sử dụng để biểu diễn sự biến động theo thời gian từng tháng. Đây là ví dụ kết quả tháng 1:

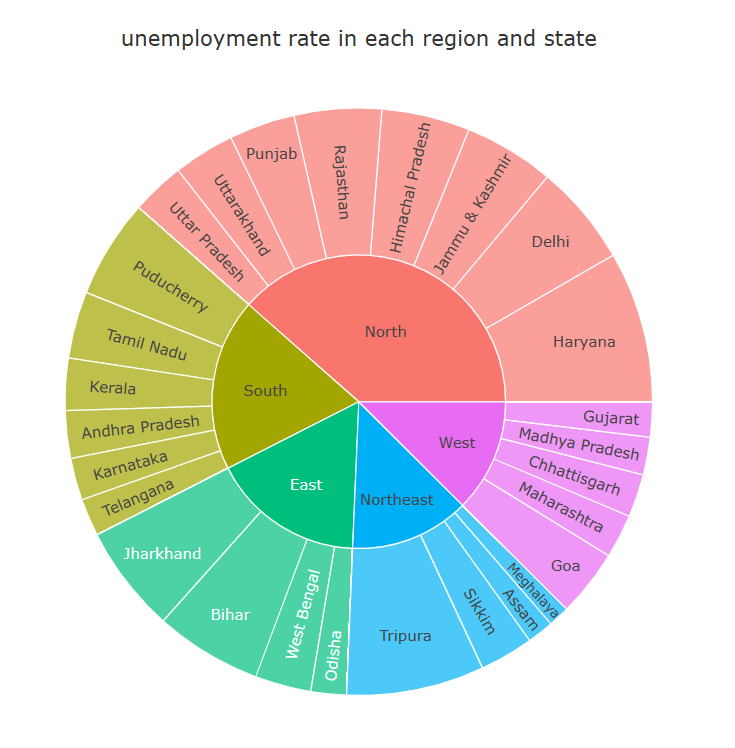


Dựa vào biểu đồ ta thấy rằng hầu hết từ tháng 1 đến tháng 10 thì khu vực North luôn giữ tỉ lệ thất nghiệp cao nhất, các khu vực còn lại có sự thay đổi nhưng không đáng kể. Nhưng có sự thay đổi lớn về tỉ lệ thất nhiệp của South và East vào tháng 4 đã vượt qua North trở thành khu vực có tỉ lệ thất nghiệp cao nhất và nhì. Hai tháng 4 và 5 ghi nhận sự tăng cao tỉ lệ của tất cả các vùng nhưng ngay các tháng sau đó tỉ lệ giảm dần và duy trì tình hình như ba tháng đầu.

**3.6. Biểu đồ Sunburst**

1. Liệu có sự chênh lệch lớn về tỷ lệ thất nghiệp giữa các bang trong cùng một khu vực không?

|  |
| --- |
| fig = px.sunburst(unemplo, path=['Region','States'], values='Estimated Unemployment Rate',  color\_continuous\_scale='Plasma',title= 'unemployment rate in each region and state',  height=650,template='ggplot2')  fig.show() |



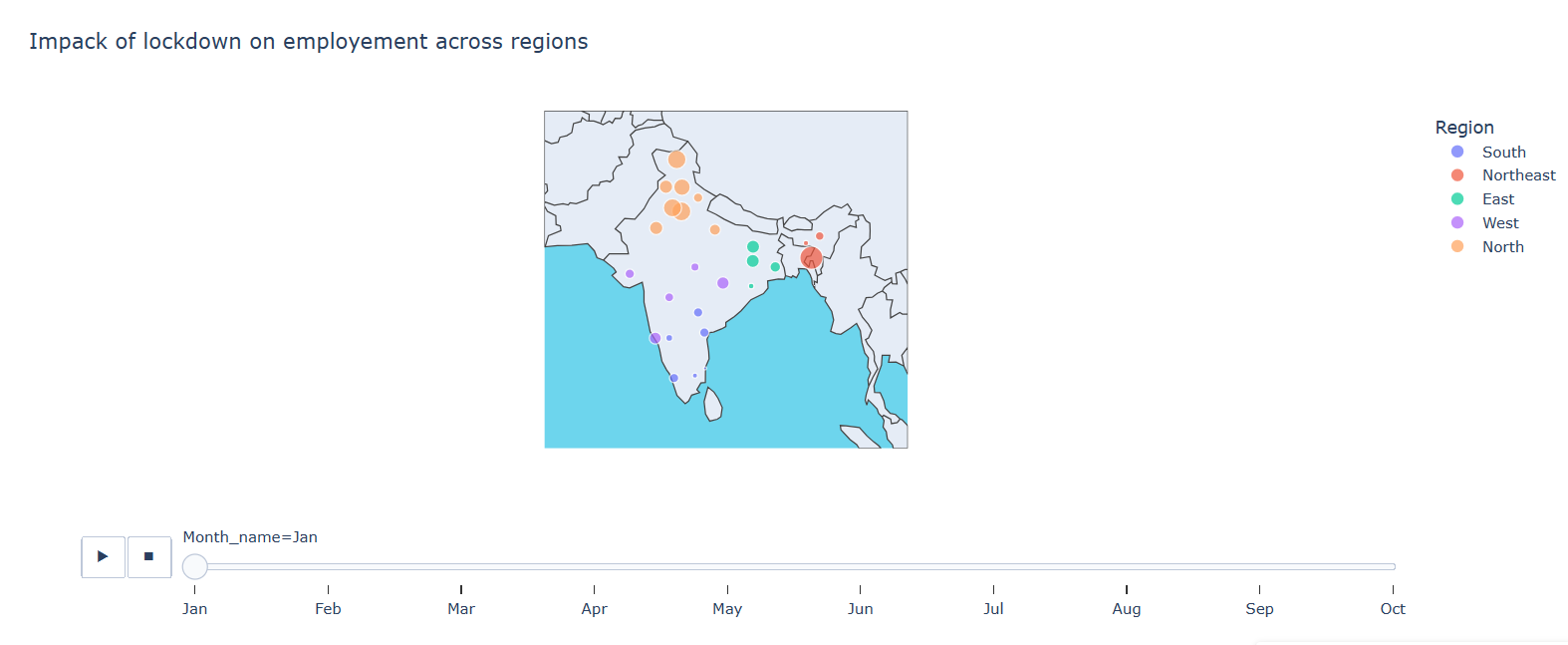
Nhìn vào biểu đồ ta có thể thấy rằng tỉ lệ thất nghiệp của hầu hết các bang trong mỗi vùng miền tương đồng nhau. Trong đó, North là miền có tỉ lệ thất nghiệp cao nhất và tại mỗi miền đều có một bang có tỉ lệ thất nghiệp vượt trội so với các bang còn lại trong vùng với North là bang Haryana, West là Goa, Northeast là Tripura, East là Jharkhand và Bihar, South là bang Puducherry. Điều này chứng tỏ tại mỗi vùng miền luôn tồn tại một bang có thể thiếu hụt cơ hội việc làm hoặc là có tỉ lệ người nghèo đói và nhiều khu ổ chuột hoặc cũng có thể số lượng việc làm không đáp ứng được so với số lượng dân lao động nhất là với đất nước đông dân như Ấn Độ. Những điều này dẫn tới việc không có sự ổn định về việc làm nói riêng và tình hình kinh tế nói chung gây ra chênh lệch lớn về tỉ lệ.

**3.7. Biểu đồ map Scatter geo**

7. Khu vực nào thường xuyên gây ra sự chênh lệch lớn nhất về tỷ lệ thất nghiệp trong mỗi khu vực qua các tháng?

|  |
| --- |
| fig = px.scatter\_geo(df,'longitude', 'latitude', color="Region",  hover\_name="States", size="Estimated Unemployment Rate",  animation\_frame="Month\_name",scope='asia',template='plotly',title='Impack of lockdown on employement across regions')  fig.layout.updatemenus[0].buttons[0].args[1]["frame"]["duration"] = 2000  fig.update\_geos(lataxis\_range=[5,35], lonaxis\_range=[65, 100],oceancolor="#6dd5ed",  showocean=True)  fig.show() |

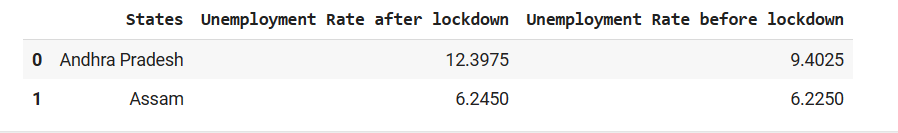
Tham số animation\_frame = 'Month\_name' được nhóm sử dụng để biểu diễn sự biến động theo thời gian từng tháng. Đây là ví dụ kết quả tháng 1:



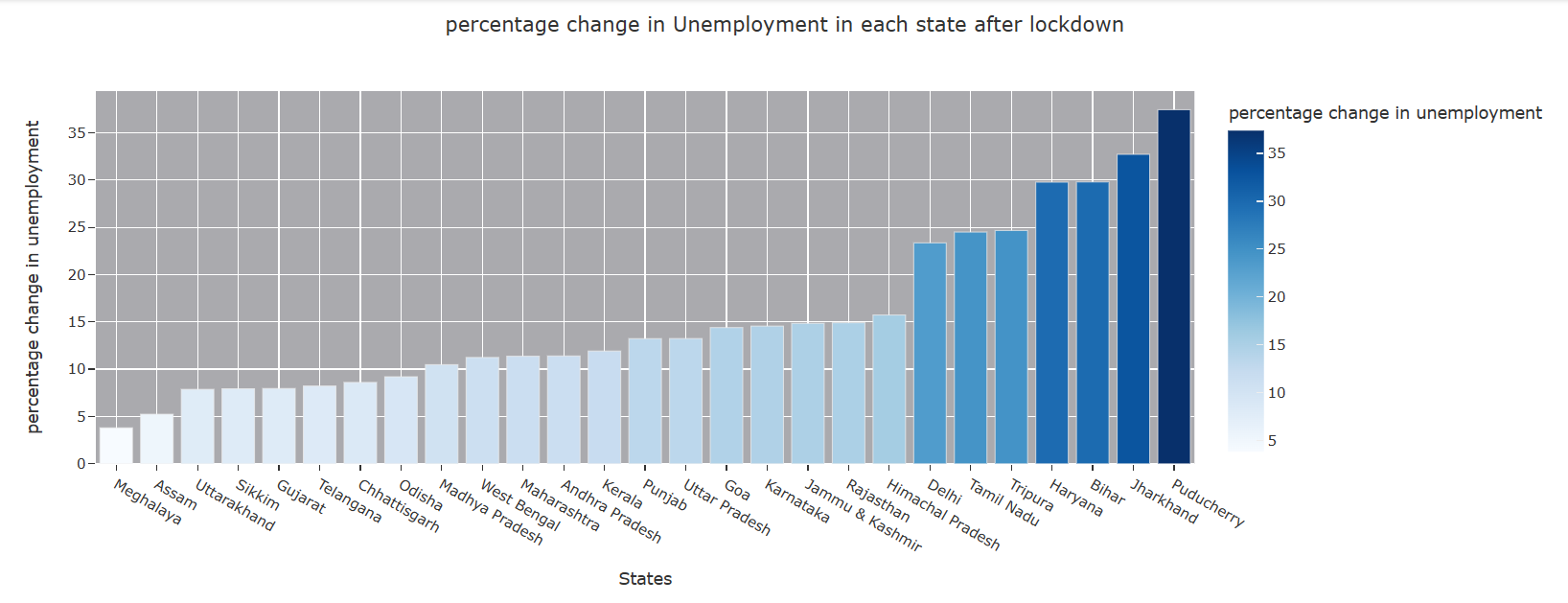
**3.8. Biểu đồ cột**

1. Bang nào có sự thay đổi lớn nhất và có đáng kể so với các bang khác sau biện pháp phong tỏa?

|  |
| --- |
| lock = df[(df['Month\_int'] >= 4) & (df['Month\_int'] <=7)]  bf\_lock = df[(df['Month\_int'] >= 1) & (df['Month\_int'] <=4)]  g\_lock = lock.groupby('States')['Estimated Unemployment Rate'].mean().reset\_index()  g\_bf\_lock = bf\_lock.groupby('States')['Estimated Unemployment Rate'].mean().reset\_index()  g\_lock['Unemployment Rate before lockdown'] = g\_bf\_lock['Estimated Unemployment Rate']  g\_lock.columns = ['States','Unemployment Rate after lockdown','Unemployment Rate before lockdown']  g\_lock.head(2) |



|  |
| --- |
| # percentage change in unemployment rate  g\_lock['percentage change in unemployment'] = round(g\_lock['Unemployment Rate after lockdown'] - g\_lock['Unemployment Rate before lockdown']/g\_lock['Unemployment Rate before lockdown'],2)  plot\_per = g\_lock.sort\_values('percentage change in unemployment')  # percentage change in unemployment after lockdown\  fig = px.bar(plot\_per, x='States', y='percentage change in unemployment', color='percentage change in unemployment',  title='percentage change in Unemployment in each state after lockdown', template='ggplot2',  color\_continuous\_scale='blues')  fig.update\_layout(  plot\_bgcolor='rgb(170, 170, 174)', # đổi màu nền của biểu đồ thành màu trắng  paper\_bgcolor='white' # đổi màu nền của giấy (tức là vùng ngoài biểu đồ) thành màu trắng  ) |



Dựa vào biểu đồ, sau biện pháp phong tỏa thì hai bang Puducherry và Jharkhand chịu ảnh hưởng lớn nhất với sự chênh lệch trước và sau biện pháp lockdown là trên 30%.

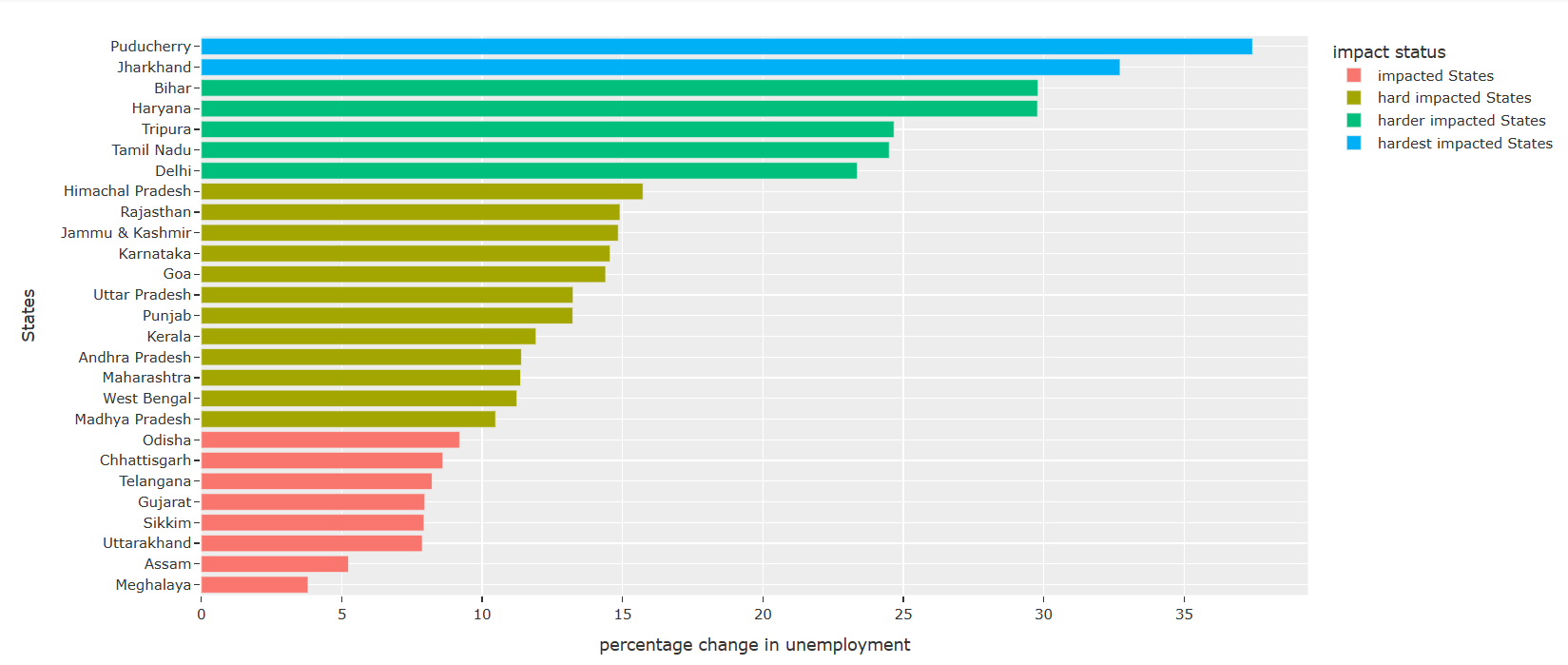
1. Nhận xét về ảnh hưởng trong thời kỳ giãn cách xã hội đối với các bang về vấn đề việc làm?

Để trả lời cho câu hỏi này nhóm có xây dựng hàm để sắp xếp và phân loại mức độ ảnh hưởng của biện pháp phong tỏa:

|  |
| --- |
| def sort\_impact(x):  if x <= 10:  return 'impacted States'  elif x <= 20:  return 'hard impacted States'  elif x <= 30:  return 'harder impacted States'  elif x <= 40:  return 'hardest impacted States'  return x |

Vẽ biểu đồ:

|  |
| --- |
| plot\_per['impact status'] = plot\_per['percentage change in unemployment'].apply(lambda x:sort\_impact(x))  fig = px.bar(plot\_per, y='States',x='percentage change in unemployment',color='impact status',  title='Impact of lockdown on employment across states',template='ggplot2',height=650)  fig.show() |



Từ biểu đồ trên ta thấy rằng:

Chịu ảnh hưởng nặng nhất gồm các bang: Puducherry, Jharkhand

Chịu ảnh hưởng nặng hơn gồm các bang: Bihar, Haryana, Tripura,…

Chịu ảnh hưởng nặng gồm các bang: Himachal Pradesh, Rajasthan, Goa, Punjab,…

Các bang còn lại có chịu ảnh hưởng.

# **KẾT LUẬN**

Một lần nữa nhóm em xin chân thành cảm ơn thầy …. đã tận tình hướng dẫn chúng em hoàn thành bài tập lớn này một cách tốt nhất.

Sau khi hoàn thành đề tài, các thành viên trong nhóm đã hiểu thêm về các biểu đồ được sử dụng, biết cách để áp dụng các biểu đồ để trực quan và từ đó hiểu thêm về bộ dữ liệu. Nhóm rút ra được:

1. **Hiểu dữ liệu**: Trực quan hóa giúp hiểu rõ hơn về cấu trúc, mẫu mực, và mối quan hệ trong dữ liệu. Bằng cách nhìn vào biểu đồ, người xem có thể dễ dàng nhận biết xu hướng, điểm nổi bật và ngoại lệ trong dữ liệu.
2. **Truyền đạt thông tin**: Trực quan hóa là một công cụ mạnh mẽ để truyền đạt thông tin một cách hiệu quả. Các biểu đồ và đồ thị giúp làm cho dữ liệu phức tạp trở nên dễ hiểu và hấp dẫn hơn cho người xem.
3. **Phát hiện xu hướng và mối quan hệ**: Bằng cách thể hiện dữ liệu dưới dạng biểu đồ, chúng ta có thể dễ dàng nhận ra các xu hướng, mẫu mực và mối quan hệ giữa các biến. Điều này giúp trong việc phân tích và dự đoán các xu hướng trong tương lai.
4. **Hỗ trợ ra quyết định**: Trực quan hóa dữ liệu cung cấp một cách để khám phá và đánh giá các phương pháp và kết quả khác nhau. Các biểu đồ và biểu đồ có thể giúp trong việc đưa ra quyết định thông minh dựa trên sự phân tích dữ liệu.
5. **Tạo sự tương tác**: Trực quan hóa dữ liệu có thể được sử dụng để tạo ra các công cụ tương tác cho người dùng, cho phép họ tương tác trực tiếp với dữ liệu và khám phá thông tin theo cách riêng của họ.

Tóm lại, trực quan hóa dữ liệu là một phương tiện quan trọng trong việc khám phá, hiểu và truyền đạt thông tin từ dữ liệu số một cách dễ dàng và hiệu quả.

**TÀI LIỆU THAM KHẢO**

[2] <https://plotly.com/python/>

[3] <https://www.w3schools.com/python/pandas/default.asp>