بنام خدا
گزارش پروژه داده کاوی
علی دینداماال ۹۷۷۳۱۱۸
۱۴۰۰-۱۴۰۱

فهرست مطالب

خلاصه:

چند سالی است که بیماری کووید ۱۹ مشکلات فراوانی را برای مردم سراسر جهان به وجود آورده. با بررسی داده های مربوط به شیوع این بیماری در سال ۲۰۲۰ میتوان کاستی ها و قدم های اشتباه کشور های متفاوت را بررسی کرد و به طور کلی از میزان آمادگی آنها تا نحوه ی مقابله با آن اطلاعاتی را بدست آورد. قصد بر این است که در این نوشته به سوالاتی پاسخ داده شود و اطلاعات بدست آمده از این دیتاست با استفاده از کتاب خانه هایی از زبان python مصور سازی شود. سپس تلاش میشود تا از روش هایی مثل خوشه بندی سعی شود تا باتوجه به هدفمان کشور ها را خوشه بندی کنیم.

دیتاست ها:

در ابتدا باید با مجموعه داده های مد نظر آشنا شویم. برای کاوش داده از دو دیتاست استفاده میشود که مرتبط به یکدیگر هستند.

اولین دیتاست با نام country_wise_latest شامل مواردی مثل نرخ مرگ و میر و موارد این چنینی برای کشور های مختلف که براثر کووید ۱۹ بوده میشود و اگر به عنوان یک دیتابیس به آن نگاه کنیم، کشور ها در جایگاه کلید اصلی قرار دارند. این دیتاست شامل ۱۵ ویژگی میشود که اسم و نوع آنها به صورت زیر است:

#	***Column***	***Non-Null Count***	***Dtype***
0	Country/Region	187 non-null	object
1	Confirmed	187 non-null	int64
2	Deaths	187 non-null	int64
3	Recovered	187 non-null	int64
4	Active	187 non-null	int64
5	New cases	187 non-null	int64
6	New deaths	187 non-null	int64
7	New recovered	187 non-null	int64
8	Deaths/100 cases	187 non-null	float64
9	Recovered/100 cases	187 non-null	float64
10	Deaths/100 recovered	187 non-null	float64
11	Confirmed last week	187 non-null	int64
12	One week change	187 non-null	int64
13	One week % change	187 non-null	float64

14	WHO Region	187 non-null	object

دومین دیتاست با نام full_grouped همانند دیتاست اول است با این تفاوت که کلید اصلی آن ترکیبی از تاریخ و کشور است. این دیتاست شامل ۱۰ ویژگی میباشد که به شرح زیر هستند:

#	***Column***	***Non-Null Count***	***Dtype***
0	Date	35156 non-null	object
1	Country/Region	35156 non-null	object
2	Confirmed	35156 non-null	int64
3	Deaths	35156 non-null	int64
4	Recovered	35156 non-null	int64
5	Active	35156 non-null	int64
6	New cases	35156 non-null	int64
7	New deaths	35156 non-null	int64
8	New recovered	35156 non-null	int64
9	WHO Region	35156 non-null	object

پاسخ به تعدادی سوال:

۱ـ پنج کشور برتر که باید در دوره های اوج شیوع ویروس کرونا به صورت کامل تعطیل میشدند کدامند؟

برای پاسخ به این سوال باید از سه فاکتور تعداد بهبود یافتگان از میان هر صد نفر، تعداد مردگان از هر صد نفر و تعداد کیس های تایید شده (که دارای بیماری هستند) از دیتاست country_wise_latest استفاده کنیم. به این صورت که ابتدا داده ها را بر اساس Death / 100 cases و بر اساس Confirmed به صورت نزولی مرتب میکنیم و در نهایت بر اساس Recovered / 100 Cases به صورت صعودی مرتب میکنیم.

قطعه کد:

```
df.sort_values(ascending=False , by=['Deaths / 100 Cases','Confirmed']).head(5).sort_values(ascending=True , by=['Recovered / 100 Cases'])
```

	Country/Region	Confirmed	Deaths	Recovered	Active	New cases	New deaths	New recovered	Deaths / 100 Cases	Recovered / 100 Cases	Deaths / 100 Recovered	Confirmed last week	1 week change	1 week % increase	WHO Region
177	United Kingdom	301708	45844	1437	254427	688	7	3	15.19	0.48	3190.26	296944	4764	1.60	Europe
16	Belgium	66428	9822	17452	39154	402	1	14	14.79	26.27	56.28	64094	2334	3.64	Europe
61	France	220352	30212	81212	108928	2551	17	267	13.71	36.86	37.20	214023	6329	2.96	Europe
184	Yemen	1691	483	833	375	10	4	36	28.56	49.26	57.98	1619	72	4.45	Eastern Mediterranean
85	Italy	246286	35112	198593	12581	168	5	147	14.26	80.64	17.68	244624	1662	0.68	Europe

که در اینجا df همان دیتا فریممان (داده هایمان) هست. و جواب این کوئری به صورت زیر برایمان نشاند داده میشود.

۲ـ کدام پنج کشور کنترل این ویروس را از دست دادند و نتوانستن درست مدیریت کنند؟

جواب این سوال وابسته به پارامتر one week % increase از دیتاست country_wise_latest هست که نشان دهنده ی درصد افزایش این بیماری در طول یک هفته چقدر بوده.

قطعه کد:

```
df.sort_values(ascending=False , by=['Confirmed','1 week % increase']).head(5)
```

	Country/Region	Confirmed	Deaths	Recovered	Active	New cases	New deaths	New recovered	Deaths / 100 Cases	Recovered / 100 Cases	Deaths / 100 Recovered	Confirmed last week	1 week change	1 week % increase	WHO Region
173	US	4290259	148011	1325804	2816444	56336	1076	27941	3.45	30.90	11.16	3834677	455582	11.88	Americas
23	Brazil	2442375	87618	1846641	508116	23284	614	33728	3.59	75.61	4.74	2118646	323729	15.28	Americas
79	India	1480073	33408	951166	495499	44457	637	33598	2.26	64.26	3.51	1155338	324735	28.11	South-East Asia
138	Russia	816680	13334	602249	201097	5607	85	3077	1.63	73.74	2.21	776212	40468	5.21	Europe
154	South Africa	452529	7067	274925	170537	7096	298	9848	1.56	60.75	2.57	373628	78901	21.12	Africa

۳ـ کدام پنج کشور کمترین تاثیر پذیری را از سوی این ویروس داشتند؟

ابتدا با استفاده از Recovered / 100 Cases کشورهایی را پیدا میکنیم که بیشترین مقدار ریکاوری را داشتند و سپس با استفاده از پارامتر Deaths و Confirmed از دیتاست country_wise_latest سعی در مرتب کردن داده ها میکنیم طوری که کمترین میزان مرگ و کیس های تایید شده را داشته باشیم.

قطعه کد:

```
df.sort_values(ascending=False , by=['Recovered / 100 Cases']).head().sort_values(ascending=True , by=['Deaths','Confirmed']).head(5)
```

	Country/Region	Confirmed	Deaths	Recovered	Active	New cases	New deaths	New recovered	Deaths / 100 Cases	Recovered / 100 Cases	Deaths / 100 Recovered	Confirmed last week	1 week change	1 week % increase	WHO Region
75	Holy See	12	0	12	0	0	0	0	0.00	100.00	0.00	12	0	0.00	Europe
49	Dominica	18	0	18	0	0	0	0	0.00	100.00	0.00	18	0	0.00	Americas
69	Grenada	23	0	23	0	0	0	0	0.00	100.00	0.00	23	0	0.00	Americas
78	Iceland	1854	10	1823	21	7	0	0	0.54	98.33	0.55	1839	15	0.82	Europe
48	Djibouti	5059	58	4977	24	9	0	11	1.15	98.38	1.17	5020	39	0.78	Eastern Mediterranean

۴ـ کدام سه کشور بیشترین نیاز را به کمک های بهداشتی دارند؟

فاکتوری که برای پاسخ به این سوال مورد استفاده قرار میگیرد شامل کیس های فعال (Active) هستند که نشاند دهنده ی وضعیت رو به بحرانی یک کشور میشود.

قطعه کد:

```
df.sort_values(ascending=False , by=['Active']).head(3)
```

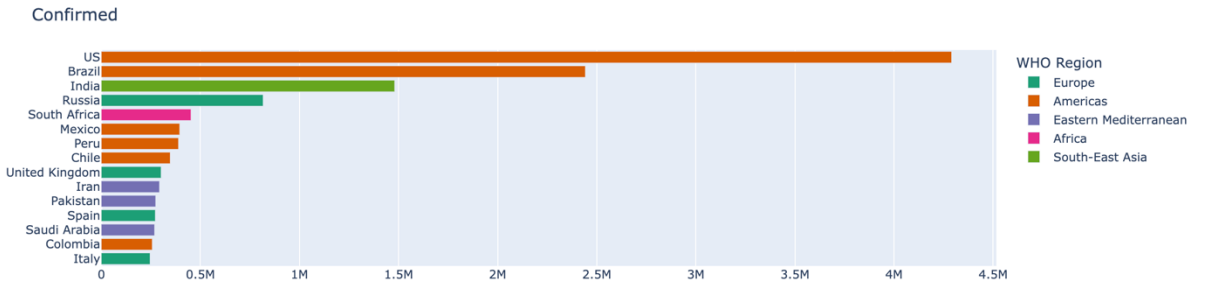
	Country/Region	Confirmed	Deaths	Recovered	Active	New cases	New deaths	New recovered	Deaths / 100 Cases	Recovered / 100 Cases	Deaths / 100 Recovered	Confirmed last week	1 week change	1 week % increase	WHO Region
173	US	4290259	148011	1325804	2816444	56336	1076	27941	3.45	30.90	11.16	3834677	455582	11.88	Americas
23	Brazil	2442375	87618	1846641	508116	23284	614	33728	3.59	75.61	4.74	2118646	323729	15.28	Americas
79	India	1480073	33408	951166	495499	44457	637	33598	2.26	64.26	3.51	1155338	324735	28.11	South-East Asia

مصور سازی داده ها:

نمودار hbar برای پانزده کشور برتر از نظر کیس های تایید شده (Confirmed) به صورت زیر میباشد:

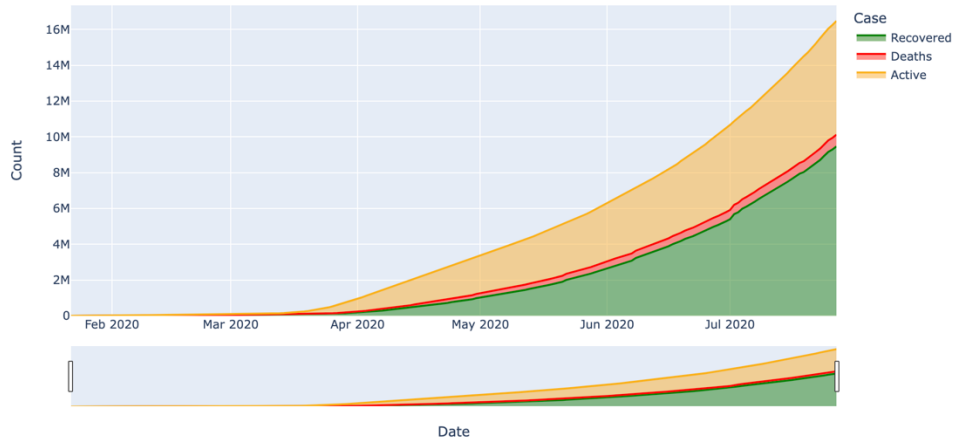
```
def plot_hbar(df, col, n, hover_data=[]):
    plot = px.bar(df.sort_values(col).tail(n),
                  x=col, y="Country/Region", color='WHO Region',
                  text=col, orientation='h', width=700, hover_data=hover_data,
                  color_discrete_sequence = px.colors.qualitative.Dark2)
    plot.update_layout(title=col, xaxis_title="", yaxis_title="",
                       yaxis_categoryorder = 'total ascending',
                       uniformtext_minsize=8, uniformtext_mode='hide')
    plot.show()
```

```
plot_hbar(df, 'Confirmed', 15)
```



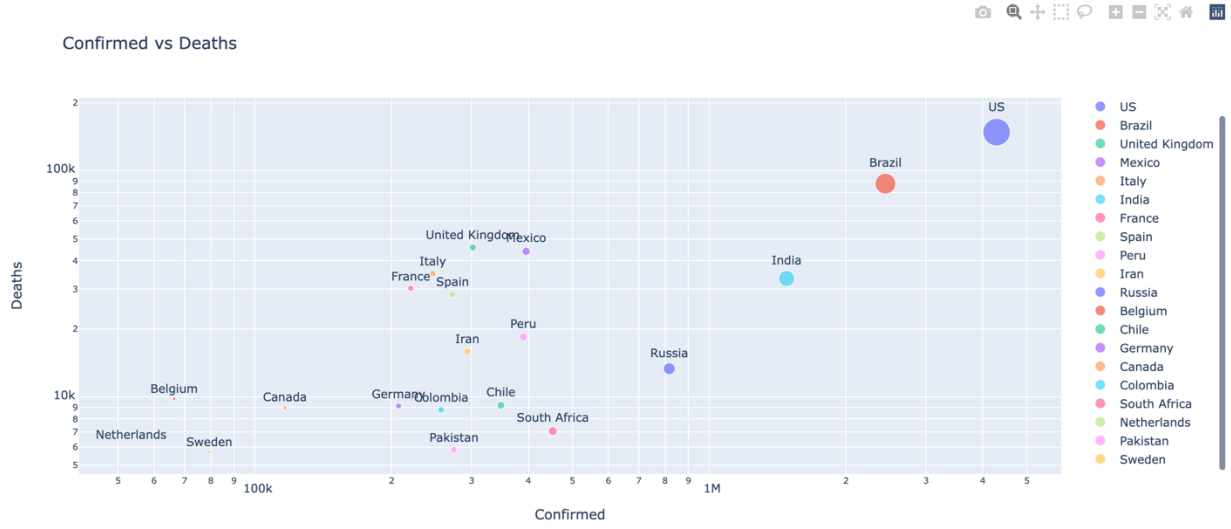
با استفاده از دیتاست full_grouped تعداد کیس های ریکاور شده(Recovered)، مرده(Deaths) و درحال حاضر فعال (Active) در طول زمان بر روی نمودار به شکل زیر هست:

```
temp = fg.groupby('Date')[['Recovered', 'Deaths', 'Active']].sum().reset_index()
temp = temp.melt(id_vars="Date", value_vars=['Recovered', 'Deaths', 'Active'],
                 var_name='Case', value_name='Count')
fig = px.area(temp, x="Date", y="Count", color='Case', height=550, width=1000,
              color_discrete_sequence = ['#008000', '#ff0000', '#fbb117'])
fig.update_layout(xaxis_rangeslider_visible=True)
fig.show()
```



استفاده از نمودار پراکندگی برای نمایش مرگ و میر(Deaths) در مقابل کیس های تایید شده(Confirmed) در بیست کشور برتر از نظر مرگ و میر (نمودار در مبنای log10 هست):

```
plot = px.scatter(df.sort_values('Deaths', ascending=False)[0:20],
                  x='Confirmed', y='Deaths', color='Country/Region', size='Confirmed',
                  height=565, text='Country/Region', log_x=True, log_y=True,
                  title='Confirmed vs Deaths')
plot.update_traces(textposition='top center')
plot.show()
```



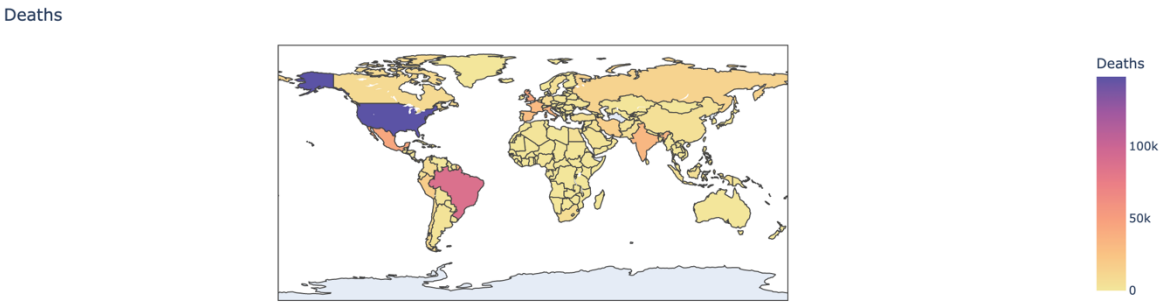
همچنین برای بررسی و مقایسه ی کشورها با استفاده از کیس های فعال(Active) میتوان از نمودار زیر استفاده کرد:

```
tm = px.treemap(df, path=["Country/Region"], values='Active', height=600, title='Active cases')
tm.data[0].textinfo = 'label+text+value'
tm.show()
```

نمودار نقشه ی کوروپلث (choropleth map) بر اساس تعداد مرگ و میر(Deaths):

```
def plot_map(df, col):
    fig = px.choropleth(df, locations="Country/Region", locationmode='country names',
                        color=col, hover_name="Country/Region",
                        title=col, hover_data=[col], color_continuous_scale='sunset')
    fig.update_layout(margin=dict(l=60, r=60, t=50, b=50))
    fig.show()

plot_map(df, 'Deaths')
```

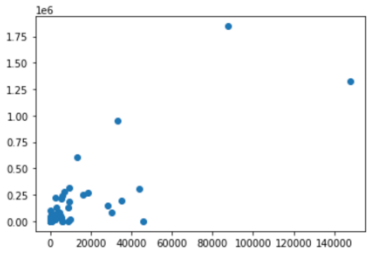


خوشه بندی:

برای خوشـــه بندی کردن براســـاس تعداد مرگ (Deaths) و تعداد بیماران بهبود یافته (Recovered) از روش K-means کافی است تعداد خوشه ها را مشخص کنیم. ابتدا نگاهی به نمودار پراکندگی بر اساس مرگ و بهبود یافتگان می اندازیم:

```
plt.scatter(df['Deaths'],df['Recovered'])
```

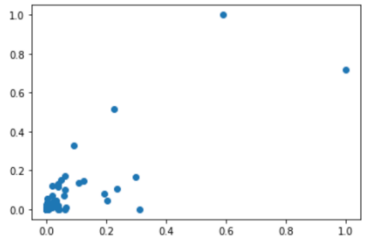
```
<matplotlib.collections.PathCollection at 0x174670a60>
```



حال سعی در نرمال سازی داده های این دو ستون میکنیم و مقدار هرکدام را در بازه ی ۰ تا ۱ معادل سازی میکنیم:

```
scaler = MinMaxScaler()
scaler.fit(df[['Deaths']])
df['Deaths'] = scaler.transform(df[['Deaths']])
scaler.fit(df[['Recovered']])
df['Recovered'] = scaler.transform(df[['Recovered']])
plt.scatter(df['Deaths'],df['Recovered'])
```

```
<matplotlib.collections.PathCollection at 0x172e271c0>
```



با توجه به نمودار تصمیم بر این است که تعداد خوشه ها را ۳ درنظر بگیریم. سپس یک ستون جدید بنام cluster میسازیم که مقدار ۰ تا ۲ را به هریک از رکورد ها اضافه میکنیم. پس به صورت زیر ادامه میدهیم:

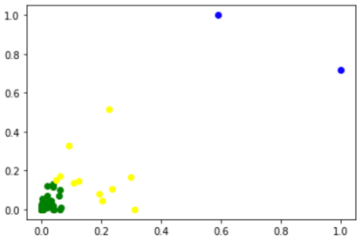
```
km = KMeans(n_clusters=3)
y_predicted = km.fit_predict(df[['Deaths','Recovered']])
df['cluster'] = y_predicted
y_predicted
```

```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 2, 0, 0, 0, 2, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       2, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       2, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
       0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0], dtype=int32)
```

بعد از ساختن ستون جدید سعی میشود تا رکوردها را باتوجه مقدار ستون جدیدشان بر روی نمودار پراکندگی با رنگ های خاص نمایش دهیم:

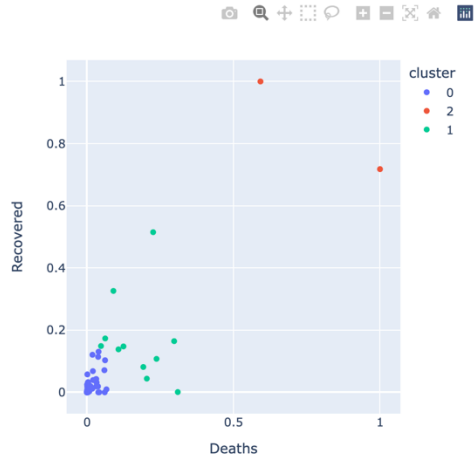
```
df0 = df[df.cluster==0]
df1 = df[df.cluster==1]
df2 = df[df.cluster==2]
plt.scatter(df0['Deaths'], df0['Recovered'], color='green')
plt.scatter(df1['Deaths'], df1['Recovered'], color='blue')
plt.scatter(df2['Deaths'], df2['Recovered'], color='yellow')
```

```
<matplotlib.collections.PathCollection at 0x1728dc070>
```



میتوانیم همین نتایج را با استفاده از یک کتاب خانه ی دیگر نیز نمایش دهیم:

```
fig = px.scatter(df, x='Deaths', y='Recovered', color='cluster', height=500, width=500)
fig.show()
```



جمع بندی:

عواملی زیادی در به وجود آمدن شـرایط بحرانی در اکثر کشـورها دخیل بودند اما با بررسـی نمودار ها میتوان متوجه این شـد که شیوع و گستردگی و مشکلات این بیماری در کشورهای با جمعیت بالا خیلی بیشتر از کشورهای دیگر است و توانایی کنترل آن نیز سـخت تر اسـت. این نمودارها و داده ها میتوانند در دسـت افراد با حرفه ی مرتبط دیدی را در اختیار آنها قرار دهند تا بتوانند از به وقوع پیوسـتن چنین شـرایطی جلوگیری کنند و اعمال لازم برای را انجام دهند و یا در صـورت بروز موج های دیگری از این بیماری درست تر از گذشته تصمیم گیری کنند تا جان انسان های کمتری در خطر بیفتد.

منابع:

دیتاست های مربوط به بیماری کرونا در سایت kaggle

نحوه ی پیاده سازی خوشه بندی به روش K-means و نحوه ی کارکردن با کتابخانه های sklearn وmatplotlib

نحوه ی کارکردن با jupyter notebook

نحوه ی کارکردن با کتابخانه ی pandas