import pandas as pd

import numpy as np

import scrapy as scr

import seaborn as sns

from matplotlib import pyplot as plt

# Importazione file csv covid19\_italy\_province.csv

file="./covid19\_italy\_province.csv"

covid\_province=pd.read\_csv(file,sep=",",index\_col="SNo")

# Per queste analisi non ci servono le colonne Country, Latitude e Longitude

province = covid\_province.drop(['Country', 'Latitude', 'Longitude'],axis=1)

# Eliminiamo le righe che contengono valori nulli

covid\_province = province.dropna()

# Importazione file csv covid19\_italy\_region.csv

file="./covid19\_italy\_region.csv"

covid\_regioni=pd.read\_csv(file,sep=",",index\_col="SNo")

# Per queste analisi non ci servono le colonne Latitude e Longitude

covid\_regioni=covid\_regioni.drop(["Latitude","Longitude"],axis=1)

# Rendere in formato data le date

covid\_regioni["Date"]=pd.to\_datetime(covid\_regioni["Date"])

covid\_regioni['Mese'] = covid\_regioni['Date'].dt.month

# Aggregazione dati per giorno

dati\_italia=covid\_regioni.groupby("Date").sum()

# Grafico temporale del 2020 in Italia relativo ai positivi giornalieri Covid-19

plt.figure(figsize=(13,6))

sns.lineplot(dati\_italia,x="Date",y="NewPositiveCases")

plt.title("Grafico temporale dei nuovi positivi giornalieri in Italia Covid-19")

plt.show()

# Aggregazione dati nuovi positivi per mese fino al 6 Dicembre 2020

covid\_regioni['mese'] = covid\_regioni['Date'].dt.to\_period('M')

somma\_per\_mese = covid\_regioni.groupby('mese')['NewPositiveCases'].sum().reset\_index()

# Progressione temporale (per mese) casi positivi Covid-19

plt.plot(somma\_per\_mese['mese'].astype(str), somma\_per\_mese['NewPositiveCases'], marker='o')

plt.xticks(rotation=45)

plt.title("Progressione temporale (per mese) casi positivi Covid-19")

plt.show()

# Trovare il numero totale di morti in Italia fino al 6 Dicembre 2020

tot\_morti=covid\_regioni.groupby(["RegionName"])["Deaths"].max().sum()

print(f"Il numero totale di morti in Italia fino al 6 Dicembre 2020 è di {tot\_morti}")

# Trovare il numero totale di positivi in Italia fino al 6 Dicembre 2020

tot\_positivi=covid\_regioni.groupby(["Country"])["NewPositiveCases"].sum().reset\_index()

total\_positive\_cases=covid\_regioni.groupby(["RegionName"])["TotalPositiveCases"].max().reset\_index()

print(f"Il numero totale di positivi in Italia fino al 6 Dicembre 2020 è di\n{tot\_positivi["NewPositiveCases"].sum()} (prendendo la colonna \'NewPositiveCases\')\n{total\_positive\_cases["TotalPositiveCases"].sum()} (prendendo la colonna \'TotalPositiveCases\')")

# Trovare il numero totale di guariti in Italia fino al 6 Dicembre 2020

tot\_guariti=covid\_regioni.groupby(["RegionName"])["Recovered"].max().sum()

print(f"Il numero totale di guariti in Italia fino al 6 Dicembre 2020 è di {tot\_guariti}")

# Variazione giornaliera positivi Covid-19

somma\_positivi\_per\_giorno = covid\_regioni.groupby('Date')['NewPositiveCases'].sum()

incremento\_giornalierio = somma\_positivi\_per\_giorno.diff().reset\_index()

print(incremento\_giornalierio)

plt.figure(figsize=(13, 6))

sns.lineplot(data=incremento\_giornalierio, x='Date', y='NewPositiveCases', label='Variazione casi positivi')

plt.title("Variazione giornaliera positivi Covid-19")

plt.show()

# Variazione mensile positivi Covid-19 fino al 6 Dicembre 2020

somma\_positivi\_per\_mese = covid\_regioni.groupby('Mese')['NewPositiveCases'].sum()

incremento\_mensile = covid\_regioni['Incremento'] = somma\_positivi\_per\_mese.diff()

print(incremento\_mensile)

# Raggruppiamo i numeri di positivi per provincia (prendendo i massimi della colonna "TotalPositiveCases" perchè quest'ultima ha valori progressivi)

max\_per\_provincia = covid\_province.groupby('ProvinceName')['TotalPositiveCases'].max().sort\_values(ascending = True)

print(max\_per\_provincia)

# Maggior e minor numero di casi di covid per provincia nel 2020

max\_assoluto\_province = max\_per\_provincia.max()

nome\_province\_max = max\_per\_provincia.idxmax()

min\_assoluto\_province = max\_per\_provincia.min()

nome\_province\_min = max\_per\_provincia.idxmin()

print(f"La provincia che ha il maggior numero di casi nel 2020 è quella di {nome\_province\_max}, con {max\_assoluto\_province} casi")

print(f"La provincia che ha il minor numero di casi nel 2020 è quella di {nome\_province\_min}, con {min\_assoluto\_province} casi")

# Media, mediana e quantili dei positivi Covid-19 per provincia

# Media

media\_provincia = max\_assoluto\_province.mean()

print(f'La media dei casi di Covid19 per provincia è {media\_provincia}')

# Mediana

mediana\_provincia = max\_per\_provincia.median()

print(f'La mediana dei positivi Covid19 per provincia è {mediana\_provincia}')

# Quartile 0.25

primo\_quartile = max\_per\_provincia.quantile(q= 0.25)

print(f"Il primo quartile (0.25) corrisponde a {primo\_quartile}")

# Quartile 0,75

terzo\_quartile = max\_per\_provincia.quantile(q= 0.75)

print(f"Il terzo quartile (0.75) corrisponde a {terzo\_quartile}")

# Grafico Boxplot della distribuzione dei positivi nelle Province

plt.figure(figsize=(7,7))

sns.boxplot(max\_per\_provincia)

plt.title("Distribuzione numeri massimi di positivi Covid-19 delle diverse province")

plt.xlabel("Province italiane")

plt.show()

# Raggruppare i nuovi positivi per regione nel 2020

somma\_positivi\_regioni=covid\_regioni.groupby(["RegionName"])["NewPositiveCases"].sum().reset\_index().set\_index("RegionName")

print(somma\_positivi\_regioni)

# La regione con meno casi nuovi nell'arco del 2020

nome\_regione\_min=somma\_positivi\_regioni.idxmin()

regione\_min=somma\_positivi\_regioni.min()

print(f"La regione con meno nuovi positivi nell'arco del 2020 è:\n{nome\_regione\_min}\nnumero di casi:\n{regione\_min}")

# Troviamo la regione con più casi nuovi nell'arco del 2020

nome\_regione\_max=somma\_positivi\_regioni.idxmax()

regione\_max=somma\_positivi\_regioni.max()

print(f"La regione con più nuovi positivi nell'arco del 2020 è:\n{nome\_regione\_max}\nnumero di casi:\n{regione\_max}")

# Numero medio di nuovi casi nell'arco del 2020 per regione

media\_regioni=somma\_positivi\_regioni.mean()

print(f"Il numero medio di nuovi casi nell'arco del 2020 è:\n{media\_regioni}")

# Regione mediana tra i nuovi casi suddivisi per ogni regione (riga scritta più per controllo che per output)

mediana\_regioni=somma\_positivi\_regioni.sort\_values(by="NewPositiveCases",ascending=True).median()

# Trovare l'indice della mediana

indice\_mediana\_regioni=(len(somma\_positivi\_regioni)-1)/2

# Ordinare la tabella somma\_positivi\_regioni in ordine crescente per numero di positivi

ordinamento\_regioni=somma\_positivi\_regioni.sort\_values(by="NewPositiveCases",ascending=True).reset\_index()

# Trovare il nome della regione per la quale il valore risulta essere la mediana

nome\_mediana\_regioni=ordinamento\_regioni.iloc[int(indice\_mediana\_regioni),0]

valore\_mediana\_regioni=ordinamento\_regioni.iloc[int(indice\_mediana\_regioni),1]

print(f"La regione mediana della tabella con i nuovi casi suddivisa per regione è:\n{nome\_mediana\_regioni} con un valore di {valore\_mediana\_regioni} nuovi casi")

# Grafico temporale giornaliero dei nuovi positivi per ogni regione nel 2020

plt.figure(figsize=(13,6))

sns.lineplot(covid\_regioni,x="Date",y="NewPositiveCases",hue="RegionName")

plt.legend(ncol=2)

plt.title("Grafico temporale dei nuovi positivi per ogni regione")

plt.show()

# Grafico a barre suddiviso per regione del numero totale dei positivi fino al 6 Dicembre 2020

somma\_positivi\_regioni=somma\_positivi\_regioni.reset\_index()

plt.figure(figsize=(10, 6))

plt.bar(somma\_positivi\_regioni["RegionName"], somma\_positivi\_regioni["NewPositiveCases"], color='g')

plt.title('Totale positivi Covid-19 suddivisi per regione')

plt.xlabel('Regione')

plt.ylabel('Totale positivi')

plt.xticks(rotation=45, ha="right")

plt.subplots\_adjust(bottom=0.20)

plt.show()

# il totale delle terapie intensive per regione

terapie\_intensive\_per\_regione=covid\_regioni.groupby("RegionName")["IntensiveCarePatients"].max()

# regione con piu terapie intensive

valore\_massimo = terapie\_intensive\_per\_regione.max()

regione\_valore\_massimo = terapie\_intensive\_per\_regione.idxmax()

print("La regione con più posti occupati in terapia intensiva è" , regione\_valore\_massimo, "con", valore\_massimo)

# Regione con meno terapie intensive

valore\_minimo= terapie\_intensive\_per\_regione.min()

regione\_valore\_minimo= terapie\_intensive\_per\_regione.idxmin()

print("La regione con meno posti occupati in terapia intensiva è", regione\_valore\_minimo,"con", valore\_minimo,"posti occupati")

media = terapie\_intensive\_per\_regione.mean()

print(f"La media dei posti occupati in terapia intensiva in Italia è {media}")

statistiche= terapie\_intensive\_per\_regione.describe()

print(statistiche)

# ordiniamo i dati e poi ci calcoliamo la mediana

terapie\_intensive\_per\_regione = terapie\_intensive\_per\_regione

median = np.median(terapie\_intensive\_per\_regione.sort\_values())

print("Mediana:", median)

# Grafico giornaliero terapie intensive nel 2020 in Italia

plt.figure(figsize=(13,6))

sns.lineplot(dati\_italia,x="Date",y="IntensiveCarePatients")

plt.title("Grafico temporale del numero giornaliero di posti occupati in terapia intensiva in Italia")

plt.show()

# Raggruppiamo i dati per regione e per mese, sommando i posti di terapia intensiva

terapia\_intensiva\_regioni\_mesi = covid\_regioni.groupby(["RegionName", "Mese"])["IntensiveCarePatients"].max().reset\_index()

# Grafico terapie intensive nel 2020 per mese e per regione

plt.figure(figsize=(13, 6))

sns.lineplot(terapia\_intensiva\_regioni\_mesi, x= 'Mese' , y ='IntensiveCarePatients', hue = 'RegionName')

plt.title("Posti occupati in terapia intensiva nel 2020 suddivisi per mese e per regione")

plt.legend(ncol=2)

plt.show()

# Rapporto morti/positivi per covid

regioni\_morti=covid\_regioni.groupby(["RegionName"])["Deaths"].max().reset\_index().set\_index("RegionName")

somma\_positivi\_regioni=somma\_positivi\_regioni.set\_index("RegionName")

rapporto\_positivi\_morti=somma\_positivi\_regioni["NewPositiveCases"]/regioni\_morti["Deaths"]

print(rapporto\_positivi\_morti)

# Rapporto percentuale tra numero positivi e abitanti per regione

# importiamo il file excel che contiene la popolazione del 2020 (preso dal istat),

# se può essere utile, il link per scaricare il file excel è questo: https://1drv.ms/x/s!AiUcVo4trr2Jh0lBTRWN3SLAim2t?e=lNj54B

popolazione = pd.read\_excel('./Popolazione\_2020\_per\_regione.xlsx')

popolazione2020= popolazione.sort\_values(by='RegionName')

# Uniamo le due tabelle somma\_positivi\_regioni e popolazione2020

tabella\_positivi\_abitanti= pd.merge(somma\_positivi\_regioni, popolazione2020, on='RegionName', how='right')

# Per calcolare l'incidenza positivi/abitanti abbiamo due modi: creare una nuova tabella oppure creare una nuova colonna nella tabella preesistente.

# Metodo 1: calcoliamo la percentuale di incidenza creando una nuova tabella

incidenza\_positivi= tabella\_positivi\_abitanti['NewPositiveCases'] / tabella\_positivi\_abitanti['Abitanti'] \*100

colonnaRegioni = tabella\_positivi\_abitanti[['RegionName']]

tabella\_incidenza\_positivi = pd.concat([colonnaRegioni,incidenza\_positivi], axis=1 )

print(tabella\_incidenza\_positivi)

# Metodo 2: aggiungiamo una colonna alla tabella unita (tabella\_positivi\_abitanti) con i valori calcolati al suo interno

tabella\_positivi\_abitanti['Rapporto Abitanti/Positivi'] = round((tabella\_positivi\_abitanti['NewPositiveCases']/tabella\_positivi\_abitanti['Abitanti']\*100), 2)

print(tabella\_positivi\_abitanti)

# Raggruppamento dati totali progressivi positivi in Italia e dei dati totali progressivi test, entrambi suddivisi per regione

test\_progressivi\_regioni= covid\_regioni.groupby(["RegionName"])[["TotalPositiveCases","TestsPerformed"]].max().reset\_index()

print(test\_progressivi\_regioni)

# Grafico totale progressivo dei tamponi fatti, suddivisi per regione

plt.figure(figsize=(13,6))

plt.bar(test\_progressivi\_regioni.sort\_values(by="TestsPerformed",ascending=False)["RegionName"],height=test\_progressivi\_regioni.sort\_values(by="TestsPerformed",ascending=False)["TestsPerformed"])

plt.title("N° di tamponi fatti suddiviso per regione")

plt.xticks(rotation=45, ha="right")

plt.subplots\_adjust(bottom=0.20)

plt.show()

# Rapporto fra positivi e test eseguiti

test\_progressivi\_regioni["Rapporto positivi/test"]=test\_progressivi\_regioni["TotalPositiveCases"]/test\_progressivi\_regioni["TestsPerformed"]

print(test\_progressivi\_regioni.sort\_values(by="TestsPerformed",ascending=False))

# Grafico rapporto tra n° di positivi e n° di tamponi fatti suddiviso per regione

plt.figure(figsize=(13,6))

plt.bar(test\_progressivi\_regioni.sort\_values(by="Rapporto positivi/test",ascending=False)["RegionName"],height=test\_progressivi\_regioni.sort\_values(by="Rapporto positivi/test",ascending=False)["Rapporto positivi/test"])

plt.title("Rapporto tra n° di positivi e n° di tamponi fatti suddiviso per regione")

plt.xticks(rotation=45, ha="right")

plt.subplots\_adjust(bottom=0.20)

plt.show()

# Vogliamo calcolare l'incidenza dei posti totali di terapia intensiva per regione con la popolazione regionale

# Incidenza delle terapie intensive sui positivi per regione

incidenza\_terapie\_intensive = round(((terapie\_intensive\_per\_regione/somma\_positivi\_regioni["NewPositiveCases"])\*100),2)

print(incidenza\_terapie\_intensive)

# Raggruppamento dati della terapia intensiva per mese e per regione

terapia\_intensiva\_regioni\_mesi = covid\_regioni.groupby(["RegionName", "Mese"])["IntensiveCarePatients"].max().reset\_index()

print(terapia\_intensiva\_regioni\_mesi)

# Creo nuova tabella da covid\_regioni per ottenere soltanto le colonne Date, RegionName e TotalHospitalizedPatients

selezione\_colonne = ['Date', 'RegionName', 'TotalHospitalizedPatients']

new\_tab = covid\_regioni[selezione\_colonne]

# La data con il maggior numero di persone ricoverate in ospedale per ogni regione

indice\_max\_ospedalizzati = new\_tab.groupby("RegionName")["TotalHospitalizedPatients"].idxmax()

massimo\_per\_regione = new\_tab.loc[indice\_max\_ospedalizzati]

massimo\_per\_regione = massimo\_per\_regione.sort\_values(by="Date",ascending=True).set\_index('Date')

print(massimo\_per\_regione)

# Numero totale nel 2020 di test fatti per ogni regione

test\_totali\_regioni = covid\_regioni.groupby("RegionName")["TestsPerformed"].max().sum()

print(f"In tutta Italia sono stati eseguiti fino al 6 Dicembre {test\_totali\_regioni} test")

# Creazione tabella dei test progressivi in Italia suddivisi per mese

test\_progressivi\_regioni= covid\_regioni.groupby(["RegionName","Mese"])["TestsPerformed"].max().reset\_index()

print(test\_progressivi\_regioni)

test\_progressivi= test\_progressivi\_regioni.groupby("Mese")["TestsPerformed"].sum().reset\_index()

print(test\_progressivi)

# Grafici temporali Italia combinati (ospedalizzati, terapia intensiva e quarantenati)

plt.figure(figsize=(13,6))

sns.lineplot(dati\_italia,x="Date",y="TotalHospitalizedPatients",label="Persone ricoverate")

sns.lineplot(dati\_italia,x="Date",y="IntensiveCarePatients",label="Persone in terapia intensiva")

sns.lineplot(dati\_italia,x="Date",y="HomeConfinement",label="Persone in quarantena")

plt.title("Grafici temporali combinati dati italiani")

plt.show()

#creazione tabella dei totali progressivi positivi in Italia suddivisi per regione e per mese

totali\_progressivi\_regioni= covid\_regioni.groupby(["RegionName","Mese"])["TotalPositiveCases"].max().reset\_index()

totali\_progressivi= totali\_progressivi\_regioni.groupby("Mese")["TotalPositiveCases"].sum().reset\_index()

print(totali\_progressivi)

# Grafico numero progressivo di positivi e numero progressivo dei test effettuati in Italia

plt.figure(figsize=(6, 4))

sns.lineplot(test\_progressivi, x="Mese", y="TestsPerformed", label = "Test Progressivi")

sns.lineplot(totali\_progressivi, x="Mese", y="TotalPositiveCases", label = "Totale Positivi")

plt.title("Andamento totale tamponi e totale positivi")

plt.show()