ACTIVIDAD
Limpieza de datos

Laura Valentina Aguilar Talero

Jose Fernando Garzon Suarez

11 de noviembre del 2023

Centro de Gestión de Mercados y Tecnologías de la información
Análisis y desarrollo de software

Ficha 2687350 ADSO

SCRIPT PYTHON

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

data =
pd.read_csv("C:\\Users\djgam\OneDrive\Documentos\Proyectos\python\COVID19-
JULIO2020.csv")
data.columns = data.columns.str.replace('DEPARTAMENTO ', 'DEPARTAMENTO')

#Graficas
print(data.shape)
data.info()

cols_cat=['PAIS','CIUDAD','SEXO','TIPO','ESTADO','ATENCION'
,'DEPARTAMENTO']
for col in cols_cat:
    print(f'Columna {col}: {data[col].nunique()} subniveles')

data.describe()
data.hist('EDAD')

print(data.shape)
data.drop_duplicates(inplace=True)
print(data.shape)

print(data['ESTADO'].unique())
data['ESTADO']=data['ESTADO'].str.replace('leve','Leve',regex=False)
print(data['ESTADO'].unique())

boxplot = data.boxplot(column=['EDAD'])
columns=['SEXO','ATENCION','TIPO','ESTADO']
fig,aix=plt.subplots(nrows=4, ncols=1, figsize=(8,30))
fig.subplots_adjust(hspace=1)

for i,col in enumerate(columns):
    aix[i].set_title(col)
    sns.countplot(x=col,data=data,ax=aix[i])

plt.show()

# Estado
DA={
    'NOMBRE': data['ESTADO']
}
DM_ESTADO=pd.DataFrame(DA)
```

```
DM_ESTADO.drop_duplicates(inplace=True)
DM_ESTADO['IDESTADO'] = range(1, len(DM_ESTADO) + 1)
DM_ESTADO.set_index('IDESTADO', inplace=True)
DM_ESTADO=DM_ESTADO.fillna('NA')


print (DM_ESTADO)
DM_ESTADO.to_csv('C:\python\DM_ESTADO.csv')
print('Guardado')


# Tipo
DA={
    'NOMBRE': data['TIPO']
}
data['TIPO']=data['TIPO'].str.replace('relacionado', 'Relacionado', regex=False)
data['TIPO']=data['TIPO'].str.replace('RELACIONADO', 'Relacionado', regex=False)
data['TIPO']=data['TIPO'].str.replace('En Estudio', 'En estudio', regex=False)


DM_TIPO=pd.DataFrame(DA)
DM_TIPO.drop_duplicates(inplace=True)
DM_TIPO['IDTIPO'] = range(1, len(DM_TIPO) + 1)
DM_TIPO.set_index('IDTIPO', inplace=True)
DM_TIPO=DM_TIPO.fillna('NA')


print (DM_TIPO)
DM_TIPO.to_csv('C:\python\DM_TIPO.csv')
print('Guardado')


# Sexo


data['SEXO']=data['SEXO'].str.replace('m', 'M', regex=False)
data['SEXO']=data['SEXO'].str.replace('f', 'F', regex=False)


DM_SEXO=pd.DataFrame(data['SEXO'])
DM_SEXO.drop_duplicates(inplace=True)
DM_SEXO['IDSEXO'] = range(1, len(DM_SEXO) + 1)
DM_SEXO.set_index('IDSEXO', inplace=True)
DM_SEXO=DM_SEXO.fillna('NA')


print (DM_SEXO)
```

```
DM_SEXO.to_csv('C:\python\DM_SEXO.csv')
print('Guardado')


# Atencion
data['ATENCION']=data['ATENCION'].str.replace('Hospital  UCI', 'UCI', regex=False)


DA={
    'NOMBRE': data["ATENCION"]
}
DM_ATENCION=pd.DataFrame(DA)
DM_ATENCION.drop_duplicates(inplace=True)
DM_ATENCION['IDATENCION'] = range(1, len(DM_ATENCION) + 1)
DM_ATENCION.set_index('IDATENCION', inplace=True)
DM_ATENCION=DM_ATENCION.fillna('NA')


print (DM_ATENCION)
DM_ATENCION.to_csv('C:\python\DM_ATENCION.csv')
print('Guardado')


# Departamento
DA={
    'IDDPTO': data["DIVIPOLA"]//1000,
    'NOMBRE': data["DEPARTAMENTO"]
}
DM_DEPARTAMENTO=pd.DataFrame(DA)
DM_DEPARTAMENTO.drop_duplicates(inplace=True)
DM_DEPARTAMENTO.set_index('IDDPTO', inplace=True)


print(DM_DEPARTAMENTO)
DM_DEPARTAMENTO.to_csv("C:\python\DM_DEPARTAMENTO.csv")
print('Guardado')


# Ciudad
DA={
    'IDCIUDAD': data["DIVIPOLA"],
    'NOMBRE': data["CIUDAD"],
    'IDDPTO': data["DIVIPOLA"]//1000
}
DM_CIUDAD=pd.DataFrame(DA)
DM_CIUDAD.drop_duplicates(inplace=True)
DM_CIUDAD.set_index('IDCIUDAD', inplace=True)
```

```
print(DM_CIUDAD)
DM_CIUDAD.to_csv("C:\python\DM_CIUDAD.csv")
print('Guardado')


# Pais
DA={
    'NOMBRE': data["PAIS"]
}
DM_PAIS=pd.DataFrame(DA)
DM_PAIS.drop_duplicates(inplace=True)


DM_PAIS['IDPAIS'] = range(1, len(DM_PAIS) + 1)
DM_PAIS.set_index('IDPAIS', inplace=True)



print(DM_PAIS)
DM_PAIS.to_csv("C:\python\DM_PAIS.csv")
print('Guardado')


# Fecha
DA={
    'NOMBRE': data["FECHA"]
}
DM_FECHA=pd.DataFrame(DA)
DM_FECHA.drop_duplicates(inplace=True)


DM_FECHA['IDFECHA'] = range(1, len(DM_FECHA) + 1)
DM_FECHA.set_index('IDFECHA', inplace=True)


print(DM_FECHA)
DM_FECHA.to_csv("C:\python\DM_FECHA.csv")
print('Guardado')


# Covid


data.columns = data.columns.str.replace('TIPO', 'IDTIPO')
data['IDTIPO']=data['IDTIPO'].str.replace('Importado', '1', regex=False)
```

```
data['IDTIPO']=data['IDTIPO'].str.replace('Relacionado', '1', regex=False)


data.columns = data.columns.str.replace('ESTADO', 'IDESTADO')
data.columns = data.columns.str.replace('DIVIPOLA', 'IDCIUDAD')
data['IDESTADO']=data['IDESTADO'].str.replace('Leve', '1', regex=False)
data['IDESTADO']=data['IDESTADO'].str.replace('Asintomatico', '2', regex=False)
data.columns = data.columns.str.replace('ATENCION', 'IDATENCION')
data['IDATENCION']=data['IDATENCION'].str.replace('Recuperado', '1',
regex=False)
data['IDATENCION']=data['IDATENCION'].str.replace('Fallecido', '2', regex=False)
data['IDATENCION']=data['IDATENCION'].str.replace('NA', '3', regex=False)
data['IDATENCION']=data['IDATENCION'].str.replace('Casa', '4', regex=False)
data['IDATENCION']=data['IDATENCION'].str.replace('UCI', '5', regex=False)
data['IDATENCION']=data['IDATENCION'].str.replace('Hospital', '6', regex=False)


data.set_index('ID', inplace=True)
data['DEPARTAMENTO']=data['IDCIUDAD']//1000
data.columns = data.columns.str.replace('DEPARTAMENTO', 'IDDPTO')


del(data['CIUDAD'])


print(data)
data.to_csv("C:\python\TH_COVID.csv")
print('Guardado')


# Subniveles


cols_cat=['PAIS','CIUDAD','SEXO','IDTIPO', 'IDESTADO', 'IDATENCION', 'IDDPTO']
for col in cols_cat:
    print(f'columna {col}: {data[col].nunique()} subniveles')



# Read the CSV files
data = pd.read_csv("C:\python\TH_COVID.csv")
pais = pd.read_csv("C:\python\DM_PAIS.csv")
fecha = pd.read_csv("C:\python\DM_FECHA.csv")



# Mapear los nombres de los países a sus respectivos IDs
```

```
pais_mapping = dict(zip(pais['NOMBRE'], pais['IDPAIS']))
# Mapear los nombres de los fechas a sus respectivos IDs
fecha_mapping = dict(zip(fecha['NOMBRE'], fecha['IDFECHA']))


# Reemplazar los valores en la columna 'PAIS' de TH_COVID con los IDs
correspondientes
data['PAIS'] = data['PAIS'].map(pais_mapping)
# Reemplazar los valores en la columna 'FECHA' de TH_COVID con los IDs
correspondientes
data['FECHA'] = data['FECHA'].map(fecha_mapping)



data.columns = data.columns.str.replace('PAIS', 'IDPAIS')
data.columns = data.columns.str.replace('FECHA', 'IDFECHA')


data.set_index('ID', inplace=True)
data.to_csv("C:\python\TH_COVID.csv")
print(data)
```

Script SQL

```
create database PythonClass;
use PythonClass;

create table DM_PAIS(
IDPAIS int primary key auto_increment,
NOMBRE text
);

create table DM_ATENCION(
IDATENCION int primary key auto_increment,
NOMBRE text
);

create table DM_ESTADO(
IDESTADO int primary key auto_increment,
NOMBRE text
);

create table DM_FECHA(
IDFECHA int primary key auto_increment,
NOMBRE text
);
```

```
create table DM_TIPO(
IDTIPO int primary key auto_increment,
NOMBRE text
);

create table DM_CIUDAD(
IDCIUDAD int primary key auto_increment,
NOMBRE text,
   IDDPTO int
);

create table DM_DEPARTAMENTO(
IDDPTO int primary key auto_increment,
NOMBRE text
);

CREATE TABLE covid19 (
   ID INT,
   IDFECHA INT,
   IDCIUDAD INT,
   IDDPTO INT,
   IDATENCION INT,
   EDAD INT,
   SEXO CHAR(1),
   IDTIPO INT,
   IDESTADO INT,
   IDPAIS INT
);

LOAD DATA INFILE 'C:\python\DM_PAIS.csv' INTO TABLE DM_PAIS
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
IGNORE 1 LINES
(IDPAIS, NOMBRE);

LOAD DATA INFILE 'C:\python\DM_ATENCION.csv' INTO TABLE DM_ATENCION
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
IGNORE 1 LINES
(IDATENCION, NOMBRE);

LOAD DATA INFILE 'C:\python\DM_ESTADO.csv' INTO TABLE DM_ESTADO
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
IGNORE 1 LINES
(IDESTADO, NOMBRE);
```

```
LOAD DATA INFILE 'C:\python\DM_FECHA.csv' INTO TABLE DM_FECHA
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
IGNORE 1 LINES
(IDFECHA, NOMBRE);

LOAD DATA INFILE 'C:\python\DM_TIPO.csv' INTO TABLE DM_TIPO
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
IGNORE 1 LINES
(IDTIPO, NOMBRE);

LOAD DATA INFILE 'C:\python\DM_CIUDAD.csv' INTO TABLE DM_CIUDAD
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
IGNORE 1 LINES
(IDCIUDAD, NOMBRE, IDDPTO);

LOAD DATA INFILE 'C:\python\DM_DEPARTAMENTO.csv' INTO TABLE
DM_DEPARTAMENTO
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
IGNORE 1 LINES
(IDDPTO, NOMBRE);

LOAD DATA INFILE 'C:\python\TH_COVID.csv' INTO TABLE covid19
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
IGNORE 1 LINES
(ID, FECHA, IDCIUDAD, IDDPTO, IDATENCION, EDAD, SEXO, IDTIPO,
IDESTADO, PAIS);

ALTER TABLE COVID19 ADD CONSTRAINT TIPOFK FOREIGN
KEY(IDTIPO)
REFERENCES DM_TIPO(IDTIPO);

ALTER TABLE COVID19 ADD CONSTRAINT ATENCIONFK FOREIGN
KEY(IDATENCION)
REFERENCES DM_ATENCION(IDATENCION);

ALTER TABLE COVID19 ADD CONSTRAINT ESTADOFK FOREIGN
KEY(IDESTADO)
REFERENCES DM_ESTADO(IDESTADO);

ALTER TABLE COVID19 ADD CONSTRAINT FECHAFK FOREIGN
KEY(IDFECHA)
REFERENCES DM_FECHA(IDFECHA);
```

```
ALTER TABLE COVID19 ADD CONSTRAINT PAISFK FOREIGN
KEY(IDPAIS)
REFERENCES DM_PAIS(IDPAIS);

ALTER TABLE COVID19 ADD CONSTRAINT CIUDADFK FOREIGN
KEY(IDCIUDAD)
REFERENCES DM_CIUDAD(IDCIUDAD);

ALTER TABLE DM_CIUDAD ADD CONSTRAINT DEPARTAMENTOFK FOREIGN
KEY(IDDPTO)
REFERENCES DM_DEPARTAMENTO(IDDPTO);

select * from covid19;
select * from dm_atencion;
select * from dm_ciudad;
select * from dm_departamento;
select * from dm_estado;
select * from dm_fecha;
select * from dm_pais;
select * from dm_tipo;
```

Modelo