



Universidad Central del Ecuador

FACULTAD DE INGENIERÍA Y CIENCIAS APLICADAS Sistemas De Información

Documento de Preparación de los Datos

Estudiante:

- Luis Angel Gaona Cumbicus
lagaona@uce.edu.ec
- Raul Alexander Pazos Erraez
rapazos@uce.edu.ec
- Cristian Daniel Toca Rocha
cdtoca@uce.edu.ec
- Marlon Josue Espinosa Mancero
mjespinosam@uce.edu.ec

Docente:

- PhD, Jefferson Tarcisio Beltrán Morales
jtbltran@uce.edu.ec

Asignatura: Minería de datos

Paralelo: S8-P2

Fecha: lunes 2 de julio de 2024





Generative AI-Powered Economic Impact Analysis System (GEIA)

Fecha:02/07/2024



Contenido

HOJA DE CONTROL	4
Historial de Cambios	4
Introducción.....	5
Preparación de los datos para series de tiempo interrumpidas	5
Preparación y Extracción de los Datos con el Evento “COVID-19”	8

**HOJA DE CONTROL**

Organismo	Universidad Central Del Ecuador		
Proyecto	Generative AI-Powered Economic Impact Analysis System (GEIA)		
Entregable	Documento de preparación de los datos		
Autor	Cristian Daniel Toca Rocha		
Versión/Edición	V1.0	Fecha Versión	02/07/2024
Aprobado por		Fecha Aprobación/...../.....
		N.º Total de Páginas	

Historial de Cambios

Fecha	Autor	Organización	Descripción



Introducción

Este documento detalla los procedimientos para la preparación de datos en el proyecto “GEIA”, cuyo objetivo es implementar un modelo predictivo para estimar los daños económicos en diversos sectores de Ecuador como resultado de eventos adversos.

Preparación de los datos para series de tiempo interrumpidas

En particular, se ha enfocado en el impacto del evento COVID-19 sobre las series temporales de los sectores económicos. Para llevar a cabo un análisis robusto de series temporales interrumpidas, los datos deben estructurarse adecuadamente. La estructura requerida para el análisis es la siguiente:

1. **Columna de Fecha:** Fecha correspondiente a cada dato registrado.
2. **Columna de Serie de Tiempo:** Serie de datos que se desea analizar (por ejemplo, ventas, comportamiento de deuda, indicadores económicos).
3. **Columnas de Variables de Control:** Datos adicionales que podrían influir en la serie analizada (por ejemplo, presupuesto de marketing, cambios legislativos, controles crediticios).

Durante el análisis del **COVID-19**, se observó que todos los sectores en Ecuador se vieron afectados por el evento. Sin embargo, algunos sectores no mostraron un impacto significativo y, por ende, se utilizaron como series de control para comparar con los sectores más afectados. Los sectores identificados como no afectados por el evento incluyen:



date	y	x ₁	x ₂	...	x _n

1. Administración y asistencia social
2. Administración pública

Estos sectores se mantuvieron relativamente estables durante la pandemia, permitiendo utilizarlos como series de control en el análisis de series temporales interrumpidas. Se procederá a combinar los datos de estos sectores con los datos de los sectores afectados para una evaluación más precisa del impacto del COVID-19.

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

# Ruta del archivo de datos Excel
file_name = "/content/Datos Unidos Sectores y Eventos.xlsx"

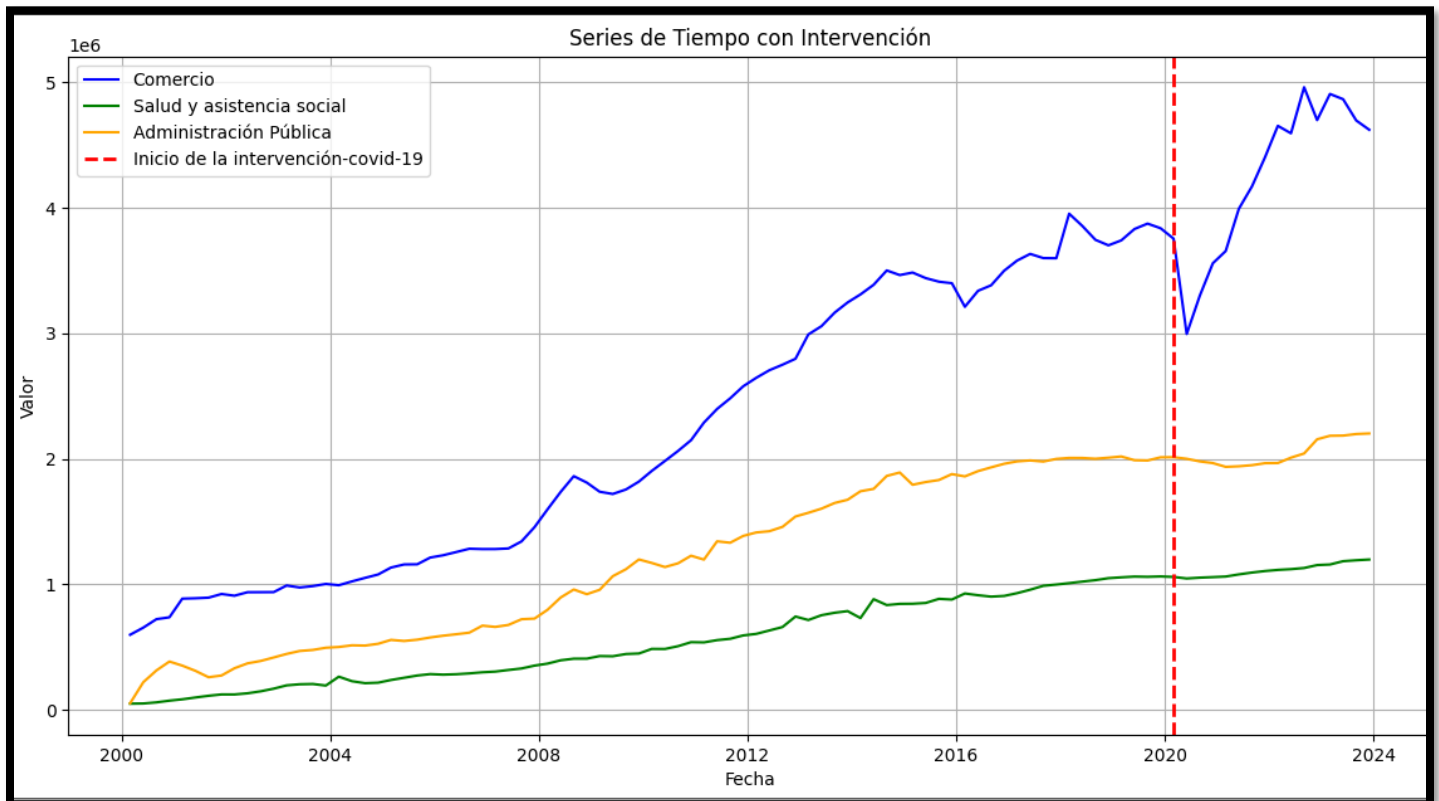
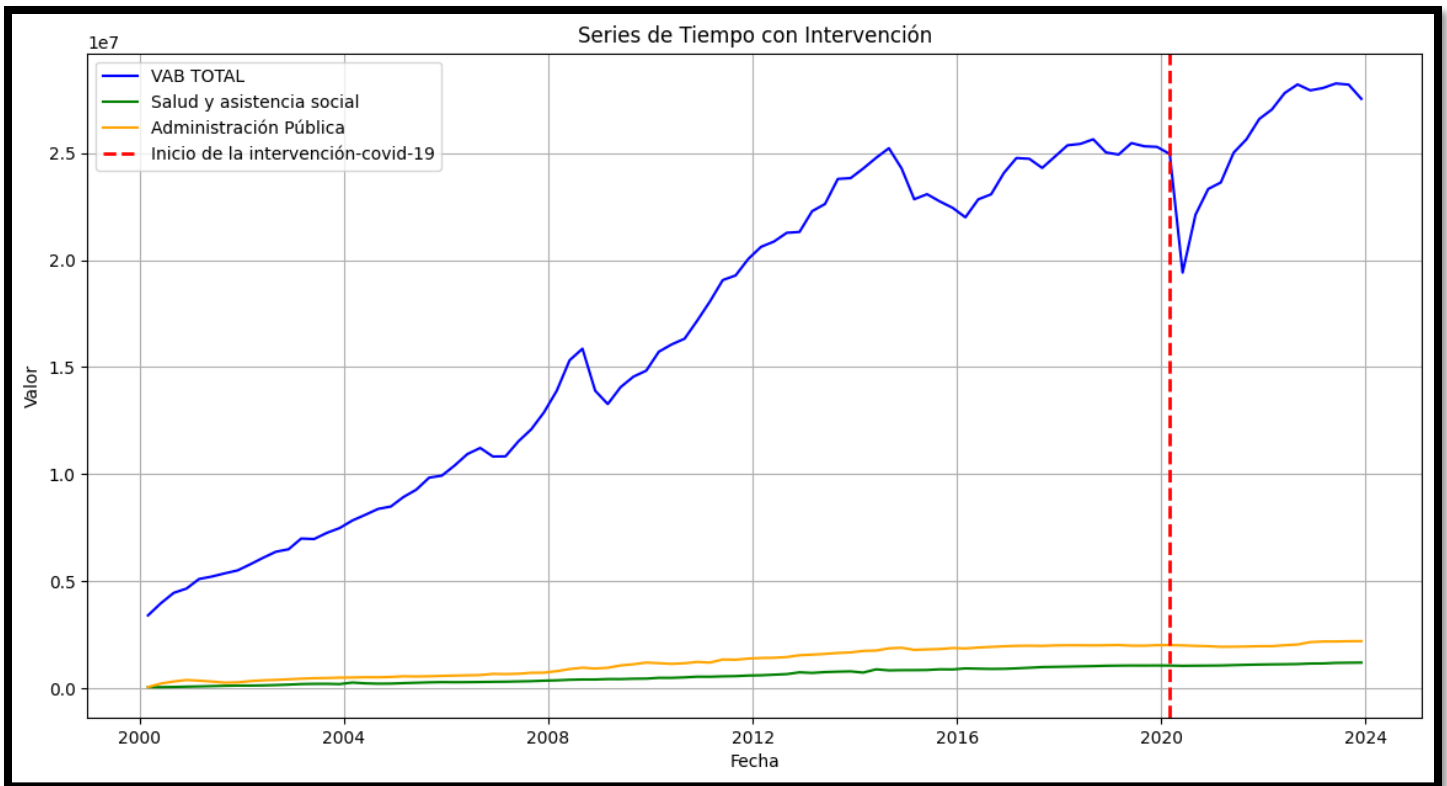
# Cargar el archivo de datos Excel y leer la hoja específica
df = pd.read_excel(file_name, sheet_name="datos_unidos")

# Función para convertir trimestre a fecha
def convertir_trimestre_a_fecha(trimestre):
    year, quarter = trimestre.split('.')
    months = {'I': 3, 'II': 6, 'III': 9, 'IV': 12}
    return pd.Timestamp(year=int(year), month=months[quarter], day=1)

# Aplicar la conversión a fecha
df['Fecha'] = df['Industrias-Trimestres'].apply(convertir_trimestre_a_fecha)
df.drop(columns=['Industrias-Trimestres'], inplace=True)

# Reordenar columnas y establecer 'Fecha' como índice
df = df[['Fecha', 'VAB TOTAL', 'Salud y asistencia social', 'Administración pública']].set_index('Fecha')

# Visualización de datos
plt.figure(figsize=(14, 7))
sns.lineplot(data=df, x=df.index, y='VAB TOTAL', label='VAB TOTAL', color='blue')
sns.lineplot(data=df, x=df.index, y='Salud y asistencia social', label='Salud y asistencia social', color='green')
sns.lineplot(data=df, x=df.index, y='Administración pública', label='Administración Pública', color='orange')
plt.axvline(pd.to_datetime('2020-03-01'), color='red', linestyle='--', linewidth=2, label='Inicio de la intervención-covid-19')
plt.title('Series de Tiempo con Intervención')
plt.xlabel('Fecha')
plt.ylabel('Valor')
plt.grid(True)
plt.legend()
plt.show()
```





Preparación y Extracción de los Datos con el Evento “COVID-19”

El siguiente código en Python ilustra el proceso de preparación y extracción de datos para el evento COVID-19. Este proceso incluye la conversión de datos trimestrales a un formato de fecha adecuado, la selección de columnas relevantes, y la generación de archivos CSV para cada sector:

```
import pandas as pd

# Ruta del archivo de datos Excel
file_name = "/content/sample_data/Datos_Unidos_Sectores_y_Eventos.xlsx"

# Cargar el archivo de datos Excel y leer la hoja específica
df = pd.read_excel(file_name, sheet_name="datos_unidos")

# Función para convertir trimestre a fecha
def convertir_trimestre_a_fecha(trimestre):
    year, quarter = trimestre.split('.')
    months = {'I': 3, 'II': 6, 'III': 9, 'IV': 12} # Mantener el ajuste original
    return pd.Timestamp(year=int(year), month=months[quarter], day=1)

# Aplicar la conversión a fecha y formatear la fecha
df['Fecha'] = df['Industrias-Trimestres'].apply(convertir_trimestre_a_fecha)
df['Fecha'] = df['Fecha'].dt.strftime('%Y-%m-%d') # Formato YYYY/MM/DD
df.drop(columns=['Industrias-Trimestres'], inplace=True)

# Definir las columnas que deben estar en el archivo
columnas = ['Agricultura ganadería y silvicultura', 'Pesca y acuicultura',
            'Explotación de minas y canteras', 'Manufactura de productos alimenticios',
            'Manufactura de productos no alimenticios', 'Refinados de petróleo',
            'Suministro de electricidad y agua', 'Construcción', 'Comercio',
            'Transporte y almacenamiento', 'Alojamiento y comidas', 'Información y comunicación',
            'Actividades financieras y de seguros', 'Actividades inmobiliarias',
            'Actividades profesionales técnicas', 'Administración pública', 'Enseñanza',
            'Salud y asistencia social', 'Arte entretenimiento y otras actividades de servicios',
            'Actividades de los Hogares como empleadores', 'VAB TOTAL',
            'Impuestos Netos Sobre los Productos', 'P.I.B.', 'Evento']

# Generar y guardar un archivo para cada columna
for columna in columnas:
    if columna not in ['Administración pública', 'Salud y asistencia social']: # Evitar duplicar estas columnas
        df_seleccionado = df[['Fecha', columna, 'Administración pública', 'Salud y asistencia social']]




















        # Guardar en formato CSV
        archivo_csv = f"/content/datos_seleccionados_{columna.replace(' ', '_')}.csv"
        df_seleccionado.to_csv(archivo_csv, index=False)
        print(f"Archivo guardado en formato CSV: {archivo_csv}")

print("Proceso completado.")
```




MINERIA DE DATOS

```
Archivo guardado en formato CSV: /content/datos_seleccionados_Agricultura_ganadería_y_silvicultura.csv
Archivo guardado en formato CSV: /content/datos_seleccionados_Pesca_y_acuicultura.csv
Archivo guardado en formato CSV: /content/datos_seleccionados_Explotación_de_minas_y_canteras.csv
Archivo guardado en formato CSV: /content/datos_seleccionados_Manufactura_de_productos_alimenticios.csv
Archivo guardado en formato CSV: /content/datos_seleccionados_Manufactura_de_productos_no_alimenticios.csv
Archivo guardado en formato CSV: /content/datos_seleccionados_Refinados_de_petroleo.csv
Archivo guardado en formato CSV: /content/datos_seleccionados_Suministro_de_electricidad_y_agua.csv
Archivo guardado en formato CSV: /content/datos_seleccionados_Construcción.csv
Archivo guardado en formato CSV: /content/datos_seleccionados_Comercio.csv
Archivo guardado en formato CSV: /content/datos_seleccionados_Transporte_y_almacenamiento.csv
Archivo guardado en formato CSV: /content/datos_seleccionados_Alojamiento_y_comidas.csv
Archivo guardado en formato CSV: /content/datos_seleccionados_Información_y_comunicación.csv
Archivo guardado en formato CSV: /content/datos_seleccionados_Actividades_financieras_y_de_seguros.csv
Archivo guardado en formato CSV: /content/datos_seleccionados_Actividades_inmobiliarias.csv
Archivo guardado en formato CSV: /content/datos_seleccionados_Actividades_profesionales_técnicas.csv
Archivo guardado en formato CSV: /content/datos_seleccionados_Enseñanza.csv
Archivo guardado en formato CSV: /content/datos_seleccionados_Arte_entretenimiento_y_otras_actividades_de_servicios.csv
Archivo guardado en formato CSV: /content/datos_seleccionados_Actividades_de_los_Hogares_como_empleadores.csv
Archivo guardado en formato CSV: /content/datos_seleccionados_VAB_TOTAL.csv
Archivo guardado en formato CSV: /content/datos_seleccionados_Impuestos_Netos_Sobre_los_Productos.csv
Archivo guardado en formato CSV: /content/datos_seleccionados_P.I.B..csv
Archivo guardado en formato CSV: /content/datos_seleccionados_Evento.csv
Proceso completado.
```

 datos_seleccionados_Actividades_de_los_Hogares_como_empleadores.csv	8/13/2024 10:26 PM	Archivo de valores...	7 KB
 datos_seleccionados_Actividades_financieras_y_de_seguros.csv	8/13/2024 10:27 PM	Archivo de valores...	7 KB
 datos_seleccionados_Actividades_inmobiliarias.csv	8/13/2024 10:27 PM	Archivo de valores...	7 KB
 datos_seleccionados_Actividades_profesionales_técnicas.csv	8/13/2024 10:27 PM	Archivo de valores...	7 KB
 datos_seleccionados_Agricultura_ganadería_y_silvicultura.csv	8/13/2024 10:27 PM	Archivo de valores...	7 KB
 datos_seleccionados_Alojamiento_y_comidas.csv	8/13/2024 10:27 PM	Archivo de valores...	7 KB
 datos_seleccionados_Arte_entretenimiento_y_otras_actividades_de_servic...	8/13/2024 10:27 PM	Archivo de valores...	7 KB
 datos_seleccionados_Comercio.csv	8/13/2024 10:27 PM	Archivo de valores...	6 KB
 datos_seleccionados_Construcción.csv	8/13/2024 10:27 PM	Archivo de valores...	6 KB
 datos_seleccionados_Enseñanza.csv	8/13/2024 10:27 PM	Archivo de valores...	6 KB
 datos_seleccionados_Explotación_de_minas_y_canteras.csv	8/13/2024 10:29 PM	Archivo de valores...	7 KB
 datos_seleccionados_Impuestos_Netos_Sobre_los_Productos.csv	8/13/2024 10:29 PM	Archivo de valores...	7 KB
 datos_seleccionados_Información_y_comunicación.csv	8/13/2024 10:29 PM	Archivo de valores...	7 KB
 datos_seleccionados_Manufactura_de_productos_alimenticios.csv	8/13/2024 10:29 PM	Archivo de valores...	7 KB
 datos_seleccionados_Manufactura_de_productos_no_alimenticios.csv	8/13/2024 10:29 PM	Archivo de valores...	7 KB
 datos_seleccionados_Pesca_y_acuicultura.csv	8/13/2024 10:29 PM	Archivo de valores...	7 KB
 datos_seleccionados_Refinados_de_petroleo.csv	8/13/2024 10:29 PM	Archivo de valores...	7 KB
 datos_seleccionados_Suministro_de_electricidad_y_agua.csv	8/13/2024 10:29 PM	Archivo de valores...	7 KB
 datos_seleccionados_Transporte_y_almacenamiento.csv	8/13/2024 10:29 PM	Archivo de valores...	7 KB



datos_seleccionados_Comercio.csv: Bloc de notas											
Archivo Edición Formato Ver Ayuda											
Fecha,Comercio,Administración pública,Salud y asistencia social											
2000-03-01,	597859.046076936,	52582.734368388185,	48905.8819342813								
2000-06-01,	654456.981353046,	219741.122147156,	50424.7603838468								
2000-09-01,	721966.596051165,	313678.354632923,	59192.9498609272								
2000-12-01,	736874.324411751,	384530.78885153594,	72897.4078209349								
2001-03-01,	885445.664093058,	351886.42183394555,	83652.7070377109								
2001-06-01,	888402.867304486,	310397.9920937321,	98143.8509879623								
2001-09-01,	893275.745072797,	259392.34599937158,	111806.503110266								
2001-12-01,	922498.52730909,	273408.240072749,	122422.938863984								
2002-03-01,	908748.671166192,	330921.4707573458,	122288.970824608								
2002-06-01,	936588.94608543,	370160.3819136823,	131552.487062408								
2002-09-01,	936974.666817642,	389088.7039959526,	147393.749199122								
2002-12-01,	937459.170780129,	416775.44333299255,	168406.792913873								
2003-03-01,	989891.177726323,	445019.601199744,	194621.161979893								
2003-06-01,	974342.881482328,	468529.2196546546,	203475.610511398								
2003-09-01,	985226.130783043,	476913.71377663803,	205676.806858077								
2003-12-01,	1002537.52485339,	494333.46536875895,	193139.420650618								
2004-03-01,	992314.889803207,	500331.5577061727,	264004.255062296								
2004-06-01,	1022889.74686178,	513688.8062717645,	227836.036380486								
2004-09-01,	1050948.46786972,	511598.9582311482,	212100.699015531								
2004-12-01,	1077894.30385557,	525887.6777908002,	216150.009541678								
2005-03-01,	1132936.45075062,	556848.0176622389,	237432.516354937								
2005-06-01,	1157745.0396688,	548217.9258739382,	255214.816163694								
2005-09-01,	1158950.34285078,	558631.3760449889,	273044.950512952								
2005-12-01,	1212753.84500886,	576128.6804187674,	284353.716968473								
2006-03-01,	1230857.56148665,	590192.8932110087,	280110.736682094								
2006-06-01,	1256640.72489596,	601911.5326640442,	283405.723122447								
2006-09-01,	1282933.90980349,	614673.4576534606,	289951.666229276								
2006-12-01,	1280103.05948736,	670310.1164716182,	298958.873966233								
2007-03-01,	1280289.25406183,	660207.9591774087,	304284.033186827								
2007-06-01,	1285025.31190279,	675998.1590197835,	317220.395578952								
2007-09-01,	1341404.7002612,	721441.0066489918,	329348.25157414								
2007-12-01,	1454865.79985063,	726535.8751537608,	351908.319660078								
2008-03-01,	1596562.76929815,	796628.6519465338,	367946.341061089								
2008-06-01,	1735431.59830658,	894364.825041997,	394362.740888724								
2008-09-01,	1860970.1360877,	958437.4306723941,	407225.289403092								
2008-12-01,	1800710.07026005,	920007.5312700601,	407450.23276000								