



Universidad
de Alcalá

**MÁSTER EN BUSINESS INTELLIGENCE AND DATA SCIENCE
EN54 - Edición 2020/21**

EFFECTOS DEL COVID 19 EN LAS ECONOMÍAS DE LOS DIFERENTES PAÍSES DE
LA UNIÓN EUROPEA.

TFM elaborado por: Angélica González González
Tutor de TFM: José Luis Llorente Perales

Madrid, Noviembre 2021

RESUMEN

Desde que en el primer trimestre del año 2020 la pandemia originada por el Covid-19, irrumpiera en Europa han pasado ya varios meses, hemos visto como en general las actividades se detuvieron casi que por completo, teniendo que reiniciar paulatinamente con las restricciones necesarias, adaptándonos a una nueva realidad. En este trabajo se indaga sobre los efectos que ha dejado esta situación en las principales economías de los países de la Unión Europea haciendo uso de algunas de las principales herramientas del big data, analizando la tasa de actividad, paro , empleo y porcentaje de población con pauta completa de vacunación.

Para esto se extraen datos públicos de organismos oficiales de la Unión Europea, asegurando el trabajo con información actualizada y verídica. Se siguen los lineamientos de la metodología CRISP-DM una de las más utilizadas en el entorno de big data y análisis de datos. Para la extracción, tratamiento y transformación de los datos se emplea Python y PySpark, y para el almacenamiento Hadoop HDFS, los resultados y visualización se presentar en gráficos generados por Plotly

Palabras clave: Covid-19, actividades económicas, mercado laboral, Big Data, Spark, Python

ABSTRACT

It has been several months since the first report of Covid cases in Europe, around the first quarter of 2020, we have seen how the life on a daily bases stopped almost entirely, and how we have manage to return on some sort of normality with the restrictions that this imply. This study quest for the economic impact of Covid-19 in European union countries, applying some of the most important tools in Big Data.

Data is extracted from official web sites of the EU in order to maintain valid and updated information. The methodology that provide the guidance for this procedure is CRISP-DM one of the most popular in the world of Big Data and Data Analysis. Python and PySpark are used for data extraction, processing and transformation, Hadoop HDFS is used for data storage, finally Plotly is used for visualization.

Key words: Covid-19, economic activities, labor force, Big Data, Spark, Python

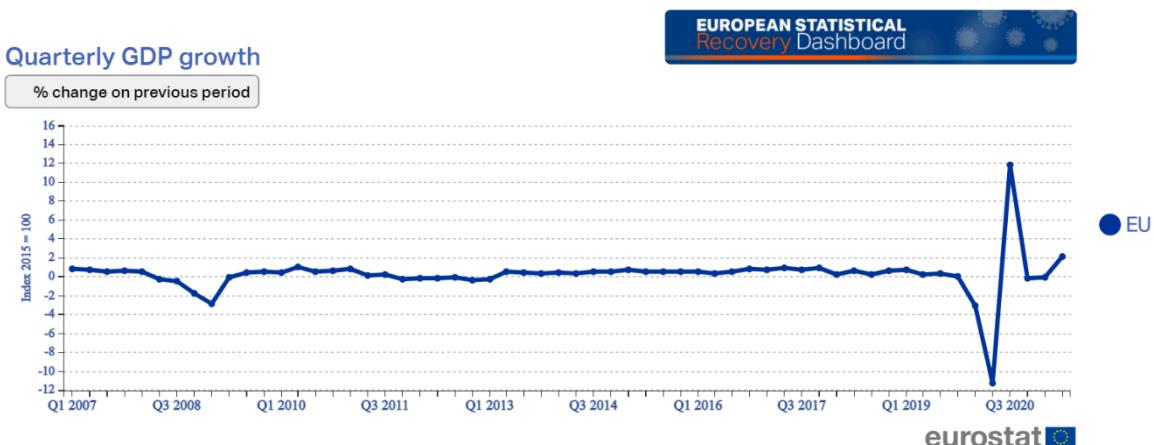
Índice

1.	Introducción	4
2.	Objetivos	6
3.	Fundamentación teórica. Estado del arte.....	7
4.	Metodología	8
5.	Desarrollo del trabajo	11
5.1.	Arquitectura	12
5.2.	Obtención de datos.....	16
5.3.	Exploración Data Sets.....	18
5.4.	Limpieza / Preparación DFs	28
5.5.	Análisis de variables	37
5.6.	Transformación de datos.....	40
5.7.	Almacenamiento	49
5.8.	Visualización y resultados	51
6.	Conclusiones	71
7.	Referencias bibliográficas	72
8.	Anexos.....	74

1. Introducción

Iniciando el 2020, la abrupta llegada de la pandemia causada por el Covid-19 “enfermedad causada por el nuevo coronavirus SARS-CoV2” (WHO, 2020) de la cual se tuvo noticia por primera vez de su existencia el 31 de diciembre de 2019 por un grupo de casos declarados en Wuhan, China, y finalmente declara como pandemia el 11 de marzo de 2020 (OMS, 2020); ha dado un vuelco a la vida como la conocíamos. Sin duda, el mayor y peor impacto lo han sufrido los sistemas sanitarios, los cuales se han visto colapsados en mayor o menor medida de un país a otro; como consecuencia de este y de las medidas de contención que los gobiernos se han visto obligados a implementar para contener la propagación del virus, se evidencian importantes y severas consecuencias económicas y sociales en todo el mundo.

Por medio de este trabajo se pretende analizar los efectos que el Covid-19 ha dejado en las economías de los 27 países miembros de la Unión Europea. En concreto, en el conjunto de la economía de la Unión Europea se registró en el segundo trimestre del 2020 una variación de -11,3% en el PIB con respecto al trimestre anterior; esta coincide con los meses en los que se registró un aumento significativo de casos positivos y muertes por Covid-19, así como la implementación de las medidas de contención más estrictas que se han tenido en el tiempo de pandemia.



Fuente: Eurostat

Figura 1: Crecimiento PIB trimestral EU porcentaje cambio con el periodo anterior

Cómo se puede ver en el grafico anterior, nos encontramos ante una situación sin precedentes, ya que este tipo de variaciones no se presentaron ni en la crisis financiera mundial del 2008. Partiendo de estos antecedentes, se hace relevante el análisis de movimientos económicos tales como los relativos al PIB, empleo y desempleo, y actividades económicas; y su correlación con el impacto que ha dejado el Covid-19 en cada uno de los países miembros de la Unión Europea y como cada uno, con sus particularidades ha sorteado esta problemática.

Para la recopilación de datos se utilizan fuentes oficiales de la Unión Europea. Para los datos relativos al Covid-19 se recurre a las bases de datos del ECDC - European Centre for Disease

Prevention and Control. Dentro de su colección de datos se dispone de data sets relativos a casos y muertes, vacunación, variantes del SARS-CoV2, hospitalización y ratios de ocupación en UCI y medidas de contención por países, estas son actualizadas en periodos semanales.

Para la recolección de datos económicos se recurre a Eurostat, la oficina estadística de la Unión Europea, cuya misión se define como: “Proporcionar estadísticas de alta calidad para Europa. Eurostat proporciona estadísticas a nivel europeo que permiten comparar países y regiones. Ofrecemos una gama completa de datos que gobiernos, empresas, el sector educativo, los periodistas y el público en general pueden usar en su trabajo y en su vida diaria” (Eurostat, 2018). En su repositorio se pueden encontrar alrededor de 8000 bases de datos socioeconómicas; a estas se puede acceder de manera interactiva desde la web o en formato tsv para descargar; también se encuentran disponibles paquetes en entornos de Python y R a para acceder a la API de Eurostat y descargar los datos directamente. Para este trabajo usé el paquete para Python.

Se define el análisis de datos como “proceso de limpieza, transformación y modelados de datos para descubrir información útil, llegando a conclusiones y soportando así el proceso de toma de decisiones” (Stephan Kudyba, 2014). A partir de las herramientas y conocimientos adquiridos durante el desarrollo del master, y con la finalidad de aplicar las mismas en un contexto real y actual, se aborda la situación socioeconómica actual de los países de la Unión Europea siguiendo las siguientes fases: Recolección, limpieza, transformación, análisis, interpretación y visualización de los datos.

2. Objetivos

Principal: Analizar a partir de bases de datos relacionales, públicas y oficiales, y con diferentes herramientas propias del big data, los efectos que la pandemia causada por el Covid-19 ha tenido sobre las economías de los países miembros de la Unión Europea.

Específicos:

- Examinar bases de datos socioeconómicas oficiales seleccionando las más actualizadas, relevantes y apropiadas para el presente estudio.
- Identificar los sectores económicos que más impacto tienen para la economía de los países de la Unión europea, para contrastar su comportamiento y variación ante la evolución de la pandemia
- Indagar sobre las tasas de empleo y desempleo en los trimestres de los años 2021, 2020 y determinar que sectores se han visto más afectados
- Comparar evolución de Covid-19 por país, vs efecto en las economías
- Contrastar evolución de vacunación con posible recuperación económica.
- Exponer análisis y resultados mediante visualización

3. Fundamentación teórica. Estado del arte

Sin duda, los esfuerzos y recursos que se han destinado a la investigación y tratamiento del virus han sido directamente proporcionales al perjuicio que este ha traído consigo. Al tiempo que científicos y personal sanitario trabajan con el objetivo de erradicar el virus y volver al “mundo prepandemia”, desde los sectores académicos, financieros económicos, tecnológicos también se han ahondado esfuerzos para investigar, analizar, predecir y mitigar las pérdidas económicas que este ha dejado a su paso.

En materia económica, instituciones como la Comisión Europea, a través del Joint Research Centre, han realizados estudios y análisis económicos a partir de los pocos meses del inicio de la pandemia; tal es el caso del artículo publicado en Agosto de 2020 “JRC analyses COVID-19 impact on economy and labor markets to help guide EU response” un análisis económico-territorial en el que se concluye que, el impacto en el mercado laboral ha sido asimétrico entre los países de la Unión europea, siendo los países que han tenido efectos más negativos aquellos que han tenido que imponer restricciones más estrictas, siendo aquellos cuya economía está sustentada en gran porcentaje en el turismo los que se han llevado el “golpe más duro”. Indica también que el PIB de media en las regiones de la UE es -6.44% con una fuerte variación de un país a otro (European Comission, 2020). Estudios como el de JRC han contribuido y sustentado, por ejemplo, la asignación de recursos del “European Recovery Package” de la Comisión Europea.

Otras investigaciones como “Impacts of the COVID-19 pandemic on EU industries” de Ecorys, encargado por el comité de Industria Investigación y Energía del Parlamento Europeo han profundizado en la necesidad de estudiar el impacto económico en las diferentes industrias exponiendo diferencias en los niveles de severidad y de impacto entre los sectores; y la necesidad de proponer medidas acorde a las necesidades de cada uno; encuentran que aquellos sectores que requieren mayor proximidad física como las industrias culturales y creativas han sido las mayores golpeadas por la crisis, mientras que los sectores farmacéuticos y tecnológicos han sido loe menos afectados. En mayo de este año preveían un crecimiento de 3.6 a 4.2% en PIB, significativamente bajo comparado con el 7.9% que se prevé en China. (Ecorys, 2021)

La economía como ciencia social, es un ente cambiante, evoluciona con el tiempo y se ve afectado por diferentes factores; cada política que toma un gobierno, una empresa o una industria tendrá inevitablemente alguna consecuencia, por esto la necesidad de mantener análisis e investigaciones actualizadas, que proporcionen descripciones, prescripciones y predicciones; así con estas generar bases sólidas para la toma de decisiones.

4. Metodología

En el campo de la ciencia de datos y a medida que esta evoluciona, se han propuesto y establecido diferentes metodologías que soportan y definen los procesos que conlleva el análisis de datos. Estas resultan importantes al momento de organizar el desarrollo de un proyecto, permiten claridad en la transmisión y progresos del mismo.

Dos que se destacan y que se sobre pesan para aplicar en el desarrollo de este proyecto son la metodología SEMMA, por sus siglas en inglés Sample, Explore, Modify, Model and Assess) y el método CRISPS-DM Cross-Industry Standard Process for Data Mining.

SEMMA desarrollada por el SAS Institute, como sus siglas lo indican contempla cinco etapas:

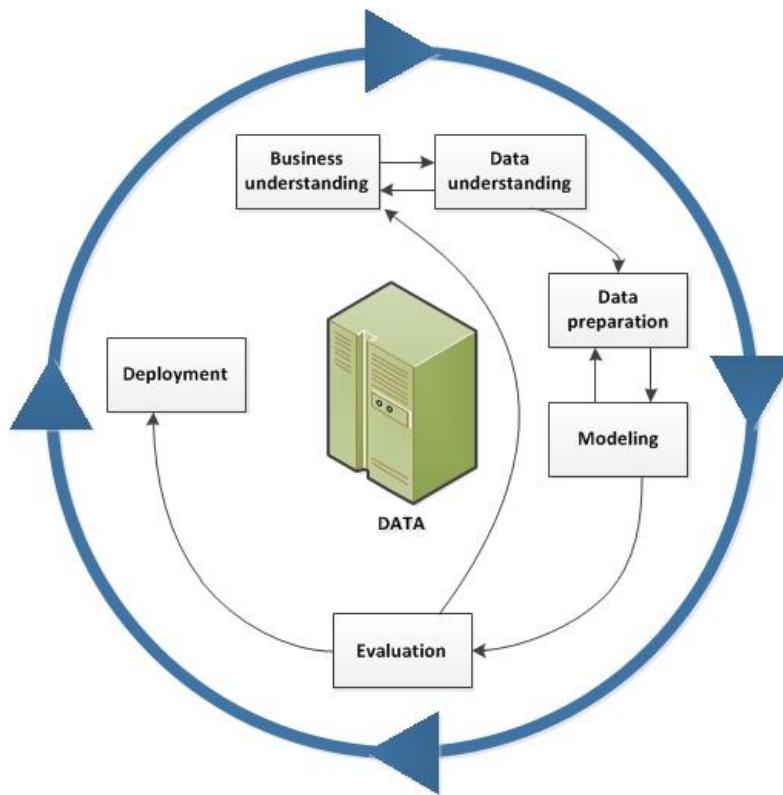
- Sample o Muestreo, en la que se extrae una muestra de los datos en una a más tablas, la muestra debe ser lo suficientemente larga que contenga información relevante y tan corta para que se permita procesar.
- Explore, se exploran los datos anticipando relaciones tendencias y anomalías
- Modify, se modifican los datos creando, seleccionando y transformando variables para enfocarlas al modelo seleccionado
- Model, en la que se determina el modelo más adecuado para obtener el resultado esperado
- Assess, se evalúa el funcionamiento del modelo de acuerdo a usabilidad y fiabilidad de los resultados del modelado.

(SAS®, 2017)

Para la planificación, desarrollo e implementación de este proyecto me apoyaré en la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), por su enfoque en el desarrollo de negocio y análisis técnico, y por ser una de las metodologías más utilizadas en el entorno de big data y análisis de datos.

Esta “proporciona una descripción normalizada del ciclo de vida de un proyecto estándar de análisis de datos, de forma análoga a como se hace en la ingeniería del software con los modelos de ciclo de vida de desarrollo de software. El modelo CRISP-DM cubre las fases de un proyecto, sus tareas respectivas, y las relaciones entre estas tareas. La metodología CRISP-DM contempla el proceso de análisis de datos como un proyecto profesional, estableciendo así un contexto mucho más rico que influye en la elaboración de los modelos.” (Sngular, 2021)

El ciclo de vida del proyecto de minería de datos se basa en seis fases:



Fuente: (IBM, 2021)

Figura 2, proceso CRISP-DM

I. Business Understanding

La fase inicial se centra en entender los objetivos y requerimientos del proyecto desde una perspectiva de negocios. En esta fase se determina:

Los objetivos de negocio: Background, criterios de éxito del negocio

Situación actual: Inventario de recursos, riesgos y contingencias, costos y beneficios

Objetivos a nivel de minería de datos: Criterios de éxito desde perspectiva técnica desde la minería de datos

Producir plan de proyecto. Seleccionar tecnologías y herramientas

II. Data Understanding

La fase de comprensión de datos inicia con la recolección de datos, procedida de tres tareas más:

Descripción de datos, se examinan propiedades de los datos tales como formato, cantidad de registros, tipos de datos.

Exploración de datos: Se indaga aún más en los datos consultas, visualizaciones, identificación de relaciones entre los datos

Verificación de la calidad del dato: que tan sucios o limpios se presentan los datos

III. Data Preparation

En esta fase se construyen los data sets finales que serán incorporados en las herramientas de modelado; se divide en cinco tareas:

1. Selección de datos: Se determina que data sets se usarán y se comentan los motivos de inclusión y/o exclusión de datos.
2. Limpieza de datos: Se corrigen, encajan o remueven valores erróneos, datos nulos
3. Construcción de datos: Se generan nuevos registros computando los existentes
4. Integración de datos: Creación de nuevos data sets combinando datos de varias fuentes
5. Dar formato a los datos: Reformatear los datos de ser necesario

IV. Modeling

En esta fase se seleccionan y aplican técnicas de modelado, se calibran los parámetros para obtener valores óptimos, se divide en cuatro tareas:

1. Selecciones técnicas de modelado: Se determinan que algoritmos se emplearán
2. Generación estrategia de verificación
3. Construcción de modelo
4. Ajustar modelo

V. Evaluation

En esta etapa se da respuesta a las preguntas: ¿la aplicación de los modelos satisface los criterios de éxito del negocio?, ¿Cuáles deberían aprobarse? Se divide en tres tareas:

1. Evaluación de resultados: Satisfacción de criterios
2. Revisión del proceso: Determinar si se han ejecutado todos los pasos correctamente, aplicar correcciones de ser necesarias
3. Determinar siguientes pasos: Se determina si se continua con el despliegue, se itera más, o se inicia un nuevo proyecto

VI. Deployment

En esta etapa se realiza el despliegue y puesta en producción. Se divide en tres tareas:

1. Plan de despliegue de modelos
2. Seguimiento y mantenimiento del plan
3. Revisión del proyecto en su globalidad

Tomado de (Chapman, y otros, 2000)

5. Desarrollo del trabajo

Siguiendo las fases de la metodología seleccionada, CRISP-DM se da inicio al proyecto entendiendo el negocio (Business Understanding), en este caso en concreto, entendiendo la problemática a analizar, revisando el back ground actual, o estado del arte, estudios previos y resultados obtenidos. Se concretan objetivos y se exploran diferentes bases de datos y herramientas, seleccionando aquellas que pudiesen contribuir de una mejor forma a la consecución de los mismos.

Una vez seleccionadas las bases de datos se procede a recolectar los datos, para temas relativos al Covid-19 se extraen del repositorio del ECDC European Centre for Disease Prevention and Control y para temas socioeconómicos se obtienen de Eurostat; fase de comprensión de datos (Data Understanding) se indaga en las propiedades de los datos recolectados, posibles relaciones entre ellos y primeras visualizaciones de los mismos, para esto se utiliza como herramienta Python, Pandas.

A continuación, se preparan los datos (Data Preparation) selección data sets, limpieza y preparación de datos se tratan los valores nulos y se evalúan y seleccionan las variables que resulten más apropiadas para el cumplimiento de los objetivos. Se almacenan como datos no elaborados

En la siguiente fase (Modeling) se inicia con la transformación de los datos, se cargan los datos limpios y con los mismo se aplican las fórmulas, uniones y modificaciones necesarias, con los resultados obtenidos se generan nuevos data sets con los que se dará paso a la siguiente fase; visualización, en esta se presentan los datos de forma legible con el objetivo de exponer resultados.

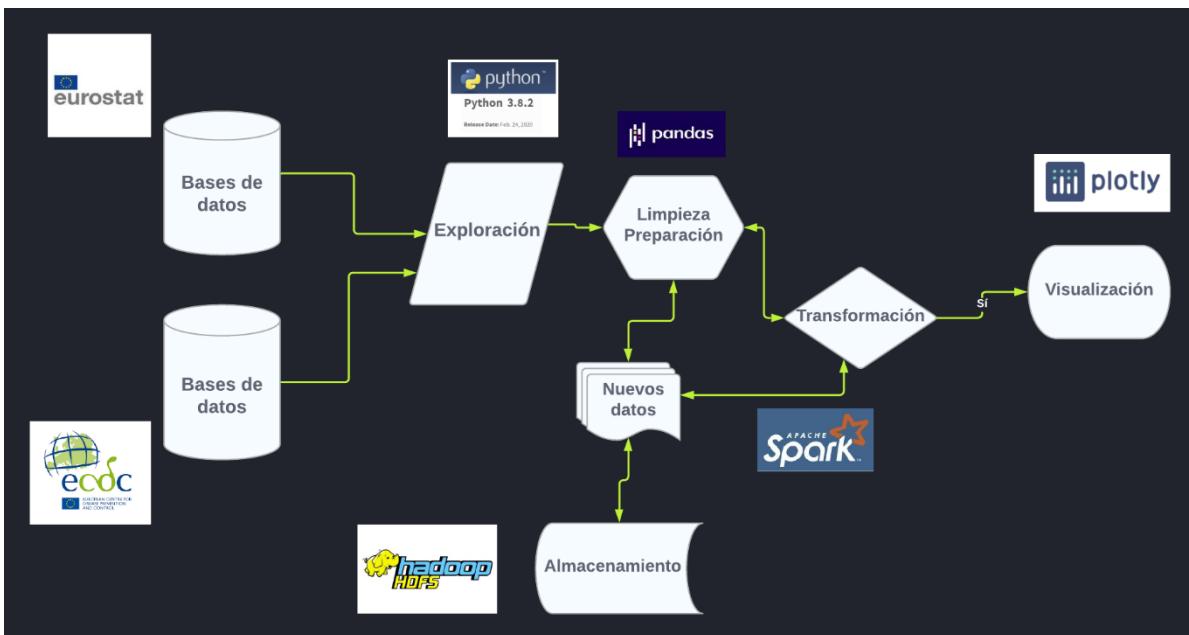


Figura 3. Flujo de trabajo

5.1. Arquitectura

A continuación, se exponen las herramientas y entornos que se emplean para el desarrollo de este proyecto; se presenta una breve descripción de cada una, su utilidad y el uso concreto para este caso en particular.

I. Python



Python 3.8.2

Release Date: Feb. 24, 2020

Figura 4. Python logo

Python, uno de los lenguajes de programación más reconocidos y utilizados, lenguaje open source de alto nivel, orientado a objetos y con semántica dinámica. Inicialmente fue

creado por Guido van Rossum como lenguaje de programación de uso general y desde su primer lanzamiento en 1991 hasta la fecha ha madurado convirtiéndose en una importante herramienta en el desarrollo del data science y Machine learning (Python.org, 2021); cuenta con un gran cantidad de librerías y su uso se ha esparcido en toda clase de dominios. En octubre de 2021 se ha posicionado en el primer lugar del ranking de TIOBE Programming Community index, indicador de popularidad de lenguajes de programación (TIOBE Software BV, 2021)

Su última versión es Python 3.10.0, lanzada el 4 de Octubre de 2021, en el desarrollo de este proyecto se emplea la versión 3.8.2, fecha de lanzamiento 24 de febrero de 2020. Se elige esta herramienta por su versatilidad, gran disponibilidad de librerías e integración con Spark. Dentro de las fases CRISP-DM se da uso principalmente en las fases data understanding y data preparation.



Figura 5. Eurostat logo

Para la recolección de datos socioeconómicos se emplea el paquete de Python Eurostat versión 0.2.3, fecha de lanzamiento 6 de Abril de 2021 herramienta para leer datos de la web de Eurostat (Cazzaniga, 2021)



Figura 6. Pandas logo

Para la recolección de datos relativos al Covid-19, descripción, exploración, limpieza de datos se hace uso de pandas, herramienta open source construida encima de Python para manipulación y análisis de datos. (pandas.pydata.org, 2021)

II. Spark



Figura 7. Spark logo

Apache Spark es un motor de análisis unificado para el procesamiento de datos a gran escala; Herramienta open source multilenguaje para la ejecución de ingeniería de datos, ciencia de datos y aprendizaje automático en máquinas de un solo nodo o clústeres

El ecosistema de Apache Spark consta de:

- Spark SQL, para el procesamiento de datos estructurados y motor de consulta SQL
- Spark Streaming, para procesar flujos de datos en tiempo real provenientes de diversas fuentes
- Mlib, biblioteca de algoritmos de Machine Learning
- GraphX, Motor de cálculo gráfico

Spark Core es el motor de ejecución subyacente de la plataforma Spark, sobre el que se construyen todas las demás funcionalidades.



Figura 8. PySpark logo

En el desarrollo de este proyecto se utiliza PySpark, interfaz para apache Spark en Python. Se elige esta herramienta por su velocidad y capacidad de procesamiento de datos a gran escala, por ser Open Source y porque es una de las herramientas predilectas en el big data. Dentro de las fases CRISP-DM se da uso principalmente en las fases data preparation y data modeling

III. HDFS (Hadoop Distributed File System)



Figura 9. HDFS logo

Es un sistema de archivos distribuido diseñado para almacenar data sets masivos con tipos de datos estructurados, semiestructurados y no estructurados. Está optimizado para almacenar grandes cantidades de datos y mantener varias copias para garantizar una alta disponibilidad y la tolerancia a fallos. HDFS es una tecnología fundamental para Big Data.

HDFS se utilizará como herramienta de almacenamiento en este proyecto, si bien el volumen de los datos que se han recolectado en este caso no son considerados como datos Big Data, se ha decidido emplear esta herramienta ya que es Open Source y por ser una de las más destacadas en el ámbito del big data y por qué se integra con Spark, ya que este no proporciona un sistema de almacenamiento de gestión de archivos, por lo que debe integrarse con otros sistemas de archivos distribuidos para funcionar, se podría otras plataformas de sistemas de datos de pago basadas en la nube como Databricks o AWS.

IV. Plotly



Figura 10. Plotly logo

Plotly es una librería gráfica open source fundada en 2013 compatible con varios lenguajes de programación como Python, R, Julia, JavaScript y MATLAB. Permite generar gráficos básicos interactivos, estadísticos, científicos entre otros. Se elige esta herramienta por su integración con otras plataformas, visualmente me parece interesante. Dentro de las fases CRISP-DM se da uso principalmente en las fases Deployment, como visualización y evaluación de los resultados del modelo.

5.2. Obtención de datos

Los datos para el desarrollo de este proyecto se recopilan de dos fuentes oficiales de la Unión Europea. Los datos relativos al Covid-19 se obtienen de la ECDC;



Figura 11. ECDC logo

El repositorio de data sets del ECDC aloja datos relativos a reportes de número de casos y muertes por Covid-19, reporte de datos de vacunación, datos sobre las variantes SARS-CoV, datos sobre el ratio de admisiones y ocupación en hospitales, datos sobre medidas de contención por países.

Estos data sets se dividen en tres categorías: Vaccination data, Daily Data y Weekle Data; los datos semanales y de vacunación se publican cada jueves, después de recolectarlos entre lunes y miércoles. Daily data arroja datos diarios relativos a casos y muertes reportados por autoridades regionales y nacionales. (ECDC, 2021) Estos datos se encuentran disponibles como archivos descargables en formatos XLSX, CSV, JSON, XML en la web de la ECDC [Download COVID-data sets.](#)

De esta fuente se utilizan dos data sets:

- ✓ Datos número de casos y muertes por COVID-19
- ✓ Datos de vacunación de Covid-19 en la Unión Europea y el Espacio Económico Europeo

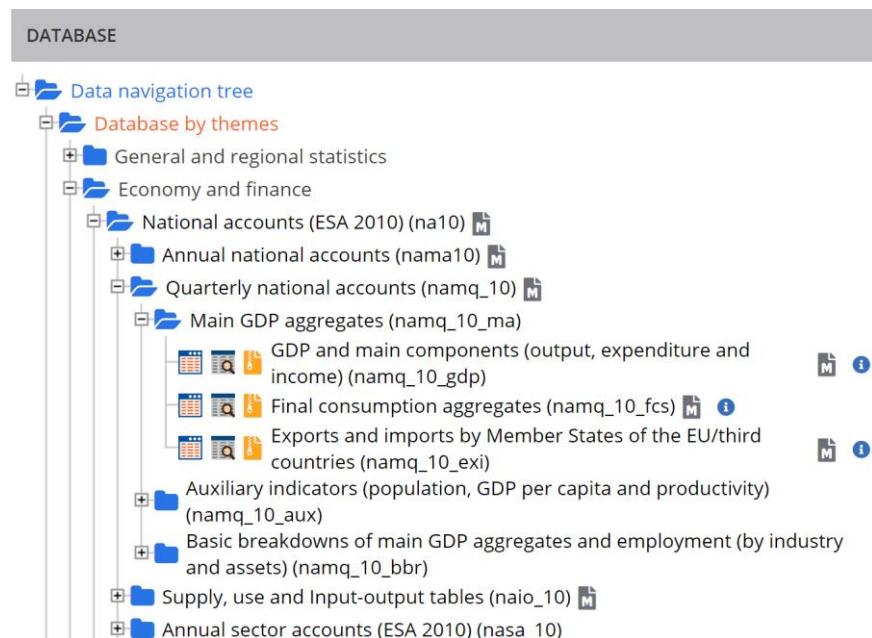
Los datos relativos a cuestiones socioeconómicas se obtienen de Eurostat;



Figura 12. Eurostat logo

Eurostat es la oficina estadística de la UE, tiene como misión proveer estadísticas y datos de gran calidad en Europa; trabaja en colaboración con institutos de estadística nacionales y otras autoridades nacionales de los estados miembros de la UE y EEE.

El repositorio de datos de Eurostat está organizado en árbol de navegación dividido en ‘Tables’ que ofrece una selección de los datos más importantes y de una manera amigable; y en ‘Database’ esta contiene el rango completo de datos publicados por Eurostat, se presenta en tablas multidimensionales con varias selecciones y formatos.



Fuente: [Eurostat](#)

Figura 13. Árbol de navegación Eurostat

Para el desarrollo de este proyecto se utilizan Data bases, desde la web se pueden descargar en formato tsv, o la opción, que he elegido utilizar en este caso, es el paquete específico para Python de Eurostat expuesto en el apartado de arquitectura de este documento.

De esta fuente se utilizan 4 data sets:

- ✓ Datos relativos al valor agregado bruto por industria
- ✓ Datos relativos a población y empleo
- ✓ Datos relativos a población por sexo, edad, ciudadanía y empleo
- ✓ Datos relativos a empleo por actividad económica

- ✓ Data Frame con los países miembros de la UE

Todos los Data Frames anteriores tienen información de otros países aparte de los 27 miembros oficiales de la Unión Europea, para filtrar los datos por solo los países miembros he creado un Data Frame con el código ISO 3166-1 alpha-2, nombre del país y el año de ingreso.

5.3. Exploración Data Sets

I. Datos número de casos y muertes por COVID-19

✓ *Data on 14-day notification rate of new COVID-19 cases and deaths*

❖ Carga de datos:

```
import pandas as pd
cases = pd.read_csv('https://opendata.ecdc.europa.eu/covid19/nation
alcasedeath/csv/data.csv')
```

❖ Resumen: Este archivo contiene información sobre notificación de nuevos casos reportados por Covid-19 por ratio de 14 días por 100000 habitantes y notificación de muertes reportadas por Covid-19 ratio de 14 días por millón de habitantes. Se actualiza semanalmente.

❖ Descripción variables:

Variable	Definición
country	Nombre del país
country_code	Código ISO de 3 letras
population	Población estadísticas Eurostat
Indicator	Casos o muertes
Weekly count	Conteo semanal
year_week	Semana reportada
rate_14_day	Ratio de 14 días por 100000 población casos o ratio de 14 días por millón de habitantes muertes
cumulative_count	Total acumulado
source	Fuente

Tabla1. Variables en DF Covid-19 cases

❖ Información general:

Dataset actualizado a: 04 de Noviembre 2021

Nº de filas / entradas: 38996

Nº de columnas / variables: 10

Tipo de datos para cada variable:

country	object
country_code	object
continent	object
population	int64
indicator	object
weekly_count	int64
year_week	object

```

rate_14_day           float64
cumulative_count     int64
source                object
dtype: object
Total valores nulos: 1596
Variables con valores nulos:
  country_code: 1152 valores nulos
  rate_14_day:   444 valores nulos

```

❖ Muestra DF

```
cases.sample(2)
```

	country	country_code	continent	population	indicator	weekly_count	year_week	rate_14_day	cumulative_count	source
38505	Yemen	YEM	Asia	29825968	cases	200	2020-28	1.163416	1465	Epidemic intelligence national data
27858	Panama	PAN	America	4314768	deaths	187	2020-29	80.884998	1096	Epidemic intelligence national data

Figura 14. Muestra DF Covid-19 cases

II. Datos de vacunación de Covid-19 en la Unión Europea y el Espacio Económico Europeo

✓ Data on COVID-19 vaccination in the EU/EEA

❖ Carga de datos:

```

import pandas as pd
Vaccine
=pd.read_csv('https://opendata.ecdc.europa.eu/covid19/vaccine_tracker/csv
/data.csv')

```

❖ Resumen: Este archivo contiene información sobre vacunación de Covid-19 en la Unión Europea y el Espacio Económico Europeo, se colectan a través del sistema europeo de vigilancia TESSy, los países miembros de la UE/EEE tienen requerido reportar indicadores básicos como el número de vacunas distribuidas por productores, cantidad de primera, segunda y no especificadas dosis administradas por grupos target a nivel nacional dos veces por semana (ECDC, 2021)

❖ Descripción variables:

Variable	Definición	Característica
YearWeekISO	Fecha en la que se ha recibido/administrado las vacunas	Año-Semana YYYY-Www
Reporting country	Código ISO 3166-1-alpha-2	Código de dos letras
Denominator	Total de población para cada grupo target por edad	Numérico
NumberDosesReceived	Numero de dosis distribuidas por productores a los países durante la semana reportada	Numérico
FirstDose	Numero de primeras dosis administradas a individuos durante la semana reportada	Numérico
FirstDoseRefused	Numero de individuos que se rehusaron a la primera dosis	Numérico
SecondDose	Numero de segundas dosis administradas a individuos durante la semana reportada	Numérico
UnknownDose	número de dosis administradas durante la semana reportada para las cuales el tipo de dosis no fue especificado	Numérico
Region	Regiones por país. Código ISO 3166	Nomenclatura NUTS1 o GAUL1
TargetGroup	Grupo target de vacunación, como mínimo los países deben reportar: "ALL" total global y "HCW" para los sanitarios	ALL = Mayores de 18 años HCW = Trabajadores sanitarios LTCFC = Usuarios de residencias Age0_4, Age5_9, Age10_14, Age15_17, Age18_24, Age25_49, Age50_59, Age60_69, Age70_79, Age<18 = niños y adolescentes Age80+ = Mayores de 80 AgeUnk = Edad desconocida 1_Age<60 = Adultos menores de 60 años 1_Age60+ = Adultos igual o mayor dde 60 años
Vaccine	Nombre de la vacuna	AZ = Vaxzevria – AstraZeneca JANSS = Ad26.COV 2.5 – Janssen COM = Comirnaty – Pfizer/BioNTech UNK = Desconocida MOD = mRNA-1273 – Moderna SPU = Sputnik V - Gamaleya BECNBG (previously CN) = Inactivated – Beijing CNBG

Tabla2. Variables en DF vaccination

❖ Información general:

```

Dataset actualizado a: 04 de Noviembre 2021
Nº de filas / entradas: 115922
Nº de columnas / variables: 12
Tipo de datos para cada variable:
YearWeekISO          object
FirstDose             int64
FirstDoseRefused      float64
SecondDose            int64
UnknownDose           int64
NumberDosesReceived   float64
Region                object
Population             int64
ReportingCountry       object
TargetGroup            object
Vaccine                object
Denominator           float64
dtype: object
Total valores nulos: 238809
Variables con valores nulos:
FirstDoseRefused: 111799 valores nulos
NumberDosesReceived: 92481 valores nulos
Denominator: 34529 valores nulos

```

❖ Muestra DF

Vaccine.sample(2)												
YearWeekISO	FirstDose	FirstDoseRefused	SecondDose	UnknownDose	NumberDosesReceived	Region	Population	ReportingCountry	TargetGroup	Vaccine	Dt	
2021-W27	0	NaN	0	0	NaN	HU	9769526	HU	AgeUNK	COM		
2021-W38	2737	NaN	12325	0	NaN	FI	5525292	FI	Age<18	COM		

Figura 15. Muestra DF Vaccination

III. Datos relativos al valor agregado bruto por industria

- ✓ Gross value added and income A*10 industry breakdowns

❖ Carga de datos:

```

import eurostat
GVA_Ind = eurostat.get_data_df('namq_10_a10')

```

- ❖ Resumen: Datos pertenecientes a las cuentas nacionales, estas se definen como conjuntos de indicadores macroeconómicos claros y coherentes que aportan una visión global de la situación económica. Eurostat publica datos anuales y trimestrales. Para este caso se trabaja con datos trimestrales.

❖ Descripción variables:

Variable	Definición	Valores
Unit	Unidad de medida	CLV05_MEUR', 'CLV05_MNAC', 'CLV10_MEUR', 'CLV10_MNAC', 'CLV15_MEUR', 'CLV15_MNAC', 'CLV_I05', 'CLV_I10', 'CLV_I15', 'CLV_PCH_PRE', 'CLV_PCH_SM', 'CON_PPCH_PRE', 'CON_PPCH_SM', 'CP_MEUR', 'CP_MNAC', 'PC_GDP', 'PC_TOT', 'PD05_EUR', 'PD05_NAC', 'PD10_EUR', 'PD10_NAC', 'PD15_EUR', 'PD15_NAC', 'PD_PCH_PRE_EUR', 'PD_PCH_PRE_NAC', 'PD_PCH_SM_EUR', 'PD_PCH_SM_NAC', 'PYP_MEUR', 'PYP_MNAC
s_adj	Ajuste estacional	CA', 'NSA', 'SA', 'SCA
nace_r2	Clasificación estadística de actividades económicas en la comunidad Europea	A', 'B-E', 'C' 'F', 'G-I', 'J', 'K', 'L', 'M_N', 'O-Q', 'R-U', 'TOTAL'
na_item	Indicador de cuentas nacionales	B1G', 'D1', 'D11', 'D12'
geo	Entidad geopolítica	AT', 'CH', 'CY', 'CZ', 'FI', 'FR', 'HU', 'IT', 'LT', 'LU', 'LV', 'TR', 'NL', 'AL', 'BA', 'BE', 'BG', 'DE', 'DK', 'EA', 'EA12', 'EA19', 'EE', 'EL', 'ES', 'EU15', 'EU27_2020', 'EU28', 'HR', 'IE', 'MK', 'MT', 'NO', 'PL', 'PT', 'RO', 'RS', 'SE', 'SI', 'SK', 'UK', 'XK', 'ME'
De 2021Q2 a 1975Q1	Trimestre en el que se reporta información	Porcentaje, miles o currency

Tabla3. Variables en DF Datos relativos al valor agregado bruto por industria

❖ Información general:

Dataset actualizado a: 29 de Octubre 2021

Nº de filas / entradas: 41210

Nº de columnas / variables: 192

Tipo de datos para cada variable:

unit object

s_adj object

nace_r2 object

na_item object

geo\time object

...

1976Q1 float64

1975Q4 float64

1975Q3 float64

1975Q2 float64

```
1975Q1      float64
Length: 192, dtype: object
Total valores nulos: 3394974
Variables con valores nulos:
 2021Q3: 39288 valores nulos
 2018Q3: 816 valores nulos
```

❖ Muestra DF

```
GVA_Ind.sample(5)
```

	unit	s_adj	nace_r2	na_item	geo\time	2021Q3	2021Q2	2021Q1	2020Q4	2020Q3	...	1977Q2	1977Q1	1976Q4
20990	CP_MNAC	SCA	F	D1	SK	NaN	424.800	434.200	415.000	437.700	...	NaN	NaN	NaN
33844	PD10_NAC	NSA	M_N	B1G	BA	NaN	118.891	118.126	118.252	118.046	...	NaN	NaN	NaN
12266	CLV_PCH_SM	SCA	TOTAL	B1G	DK	NaN	9.200	-0.500	-1.100	-1.900	...	NaN	NaN	NaN
21245	CP_MNAC	SCA	J	D1	CY	NaN	132.600	130.500	128.800	132.000	...	NaN	NaN	NaN
40339	PYP_MEUR	SCA	C	B1G	UK	NaN	NaN	NaN	NaN	49249.800	...	NaN	NaN	NaN

Figura 15. Muestra DF Datos relativos al valor agregado bruto por industria

IV. Datos relativos a población y empleo

✓ *Population and employment*

❖ Carga de datos:

```
import eurostat
empl = eurostat.get_data_df('namq_10_pe')
```

❖ Resumen:

Indicadores clave auxiliares, su uso deriva en agregados principales de PIB per cápita, productividad y coste unitario de fuerza laboral

❖ Descripción variables:

Variable	Definición	Valores
Unit	Unidad de medida	PCH_SM_PER', 'THS_PER', 'PCH_PRE_PER'
s_adj	Ajuste estacional	CA', 'NSA', 'SA', 'SCA
na_item	Indicador de cuentas nacionales	POP_NC', 'SAL_NC', 'SELF_DC', 'SAL_DC', 'EMP_NC', 'SELF_NC', 'EMP_DC'

geo	Entidad geopolítica	AT', 'CH', 'CY', 'CZ', 'FI', 'FR', 'HU', 'IT', 'LT', 'LU', 'LV', 'TR', 'NL', 'AL', 'BA', 'BE', 'BG', 'DE', 'DK', 'EA', 'EA12', 'EA19', 'EE', 'EL', 'ES', 'EU15', 'EU27_2020', 'EU28', 'HR', 'IE', 'MK', 'MT', 'NO', 'PL', 'PT', 'RO', 'RS', 'SE', 'SI', 'SK', 'UK', 'XK', 'ME'
De 2021Q2 a 1975Q1	Trimestre en el que se reporta información	Porcentaje, miles o currency

Tabla4. Variables en DF Datos relativos a población y empleo

❖ Información general:

```
Dataset actualizado a: 29 de Octubre 2021
Nº de filas / entradas: 1264
Nº de columnas / variables: 191
Tipo de datos para cada variable:
unit          object
s_adj         object
na_item        object
geo\time      object
2021Q3       float64
...
1976Q1       float64
1975Q4       float64
1975Q3       float64
1975Q2       float64
1975Q1       float64
Length: 191, dtype: object
Total valores nulos: 109075
Variables con valores nulos:
2021Q3: 1204 valores nulos
2018Q3: 66 valores nulos
```

❖ Muestra DF

```
PopEmpl.sample(5)
```

	unit	s_adj	na_item	geo\time	2021Q3	2021Q2	2021Q1	2020Q4	2020Q3	2020Q2	...	1977Q2	1977Q1	1976Q4	1976Q3	1976Q2
1081	THS_PER	SA	SELF_NC	RS	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN
800	THS_PER	NSA	EMP_DC	MT	NaN	261.16	259.29	258.66	257.20	257.51	...	NaN	NaN	NaN	NaN	NaN
62	PCH_PRE_PER	SCA	EMP_DC	EA	NaN	0.70	-0.10	0.40	1.00	-3.00	...	NaN	NaN	NaN	NaN	NaN
982	THS_PER	NSA	SELF_DC	MT	NaN	33.35	34.25	34.26	33.83	33.40	...	NaN	NaN	NaN	NaN	NaN
1006	THS_PER	NSA	SELF_NC	HR	NaN	239.36	213.31	211.14	229.74	221.46	...	NaN	NaN	NaN	NaN	NaN

5 rows × 191 columns

Figura 16. Muestra DF Datos relativos a población y empleo

V. Datos relativos a población por sexo, edad, ciudadanía y empleo

- ✓ *Population by sex, age, citizenship and labour status*

- ❖ Carga de datos:

```
import eurostat
PopByAge = eurostat.get_data_df()
```

- ❖ Resumen: Datos concernientes a situación laboral población con distinción de edad, sexo y nacionalidad
- ❖ Descripción variables:

Variable	Definición	Valores
Unit	Unidad de medida	THS = Thousand
sex	Sexo	F = Female, T = Total, M = Males
citizen	Nacionalidad	NEU28_FOR', 'NRP', 'EU28_FOR', 'STLS' 'EU27_2020_FOR', 'NEU27_2020_FOR', 'TOTAL', 'EU15_FOR', 'NEU15_FOR ', 'FOR', 'NAT'
age	Rango de edades	Y25-29', 'Y50-64', 'Y25-49', 'Y65-69', 'Y15-24', 'Y40-64', 'Y70-74', 'Y40-44', 'Y25-54', 'Y15-64', 'Y_GE15', 'Y35-39', 'Y_GE65', 'Y60-64', 'Y45-49', 'Y55-59', 'Y50-54', 'Y_GE75', 'Y20-64', 'Y15-19', 'Y25-64', 'Y20-24', 'Y55-64', 'Y15-74', 'Y40-59', 'Y25-59', 'Y50-59', 'Y_GE50', 'Y15-59', 'Y_GE25', 'Y30-34',
wstatus:	Situación laboral	POP', 'INAC', 'ACT', 'UNK', 'EMP', 'UNE'
De 2021Q2 a 1975Q1	Trimestre en el que se reporta información	Porcentaje, miles o currency

Tabla 5. Datos relativos a población por sexo, edad, ciudadanía y empleo

- ❖ Información general:

```
Dataset actualizado a: 29 de Octubre 2021
Nº de filas / entradas: 188206
Nº de columnas / variables: 100
Tipo de datos para cada variable:
unit          object
sex          object
```

```

citizen      object
age          object
wstatus      object
...
1999Q1       float64
1998Q4       float64
1998Q3       float64
1998Q2       float64
1998Q1       float64
Length: 100, dtype: object
Total valores nulos: 10289711
Variables con valores nulos:
2021Q2: 129777 valores nulos
2018Q3: 79421 valores nulos

```

❖ Muestra DF

PopByAge.sample(3)																			
	unit	sex	citizen	age	wstatus	geotime	2021Q2	2021Q1	2020Q4	2020Q3	...	2000Q2	2000Q1	1999Q4	1999Q3	1999Q2	1999Q1	1998Q4	1998Q3
139763	THS	T	EU28_FOR	Y40-64	EMP	SE	NaN	NaN	NaN	NaN	...	NaN							
139501	THS	T	EU28_FOR	Y40-44	UNE	UK	NaN	NaN	NaN	NaN	...	NaN							
11800	THS	F	EU28_FOR	Y15-39	INAC	PL	NaN	NaN	NaN	NaN	...	NaN							

Figura 17. Muestra Datos relativos a población por sexo, edad, ciudadanía y empleo

VI. Datos relativos a empleo por actividad económica

- ✓ Employment A*10 industry breakdowns

❖ Carga de datos:

```

import eurostat
emplInd = eurostat.get_data_df('namq_10_a10_e')

```

❖ Resumen: Datos con indicadores de empleo para cada una de las actividades NACE

❖ Descripción variables:

Variable	Definición	Valores
Unit	Unidad de medida	'I15_HW', 'I15_JOB', 'I15_PER', 'PCH_PRE_HW', 'PCH_PRE_JOB', 'PCH_PRE_PER', 'PCH_SM_HW', 'PCH_SM_JOB', 'PCH_SM_PER', 'PC_TOT_HW', 'PC_TOT_JOB', 'PC_TOT_PER', 'THS_HW', 'THS_JOB', 'THS_PER'

nace_r2	Clasificación estadística de actividades económicas en la comunidad Europea	A', 'B-E', 'C' 'F', 'G-I', 'J', 'K', 'L', 'M_N', 'O-Q', 'R-U', 'TOTAL'
s_adj	Ajuste estacional	CA', 'NSA', 'SA', 'SCA
na_item	Indicador de cuentas nacionales	'EMP_DC' 'SAL_DC' 'SELF_DC'
De 2021Q2 a 1975Q1	Trimestre en el que se reporta información	Porcentaje, miles o currency

Tabla 6. Datos relativos a empleo por actividad económica

❖ Información general:

Dataset actualizado a: 29 de Octubre 2021

Nº de filas / entradas: 26992

Nº de columnas / variables: 192

Tipo de datos para cada variable:

unit object

nace_r2 object

s_adj object

na_item object

geo\time object

...

1976Q1 float64

1975Q4 float64

1975Q3 float64

1975Q2 float64

1975Q1 float64

Length: 192, dtype: object

Total valores nulos: 2339542

Variables con valores nulos:

2021Q3: 25951 valores nulos

2018Q3: 1551 valores nulos

❖ Muestra DF

	unit	nace_r2	s_adj	na_item	geo\time	2021Q3	2021Q2	2021Q1	2020Q4	2020Q3	...	1977Q2	1977Q1	1976Q4	1976Q3	1976Q2	1976Q1
25099	THS_PER	F	SCA	SELF_DC	UK	NaN	NaN	NaN	NaN	852.26	...	NaN	NaN	NaN	NaN	NaN	NaN
5990	PCH_PRE_HW	A	SCA	EMP_DC	LT	NaN	2.4	-7.2	-2.7	3.30	...	NaN	NaN	NaN	NaN	NaN	NaN
24499	THS_PER	B-E	NSA	SELF_DC	IS	NaN	0.2	0.2	0.3	0.20	...	NaN	NaN	NaN	NaN	NaN	NaN

Figura 18. Datos relativos a empleo por actividad económica

5.4. Limpieza / Preparación DFs

I. Datos número de casos y muertes por COVID-19

✓ *Data on 14-day notification rate of new COVID-19 cases and deaths*

- Filtro:

Este DF contiene información de diferentes países alrededor del mundo, así como de continentes enteros, filtro los datos de los 27 países miembros de la UE. Los valores nulos en la variable *country_code* corresponden a datos para el total de un continente, por lo que al filtrar por los países objetivo de este proyecto espero no tener valores nulos para esta variable

```
# filtrar por los países miembros de La UE
from MyFunctions import EUcountries

cases = cases[cases.country.isin(EUcountries().Country)]

#Verifico que estén los 27
cases['country'].unique()

array(['Austria', 'Belgium', 'Bulgaria', 'Croatia', 'Cyprus', 'Czechia',
       'Denmark', 'Estonia', 'Finland', 'France', 'Germany', 'Greece',
       'Hungary', 'Ireland', 'Italy', 'Latvia', 'Lithuania', 'Luxembourg',
       'Malta', 'Netherlands', 'Poland', 'Portugal', 'Romania',
       'Slovakia', 'Slovenia', 'Spain', 'Sweden'], dtype=object)
```

Figura 19. Filtro UE countries

- Tratamiento valores nulos:

Después filtrar quedan 54 valores nulos en la variable *rate_14_day*, los visualizo:

```
print("Valores nulos en rate_14_day")
cases[cases.rate_14_day.isnull()].sample(54)

Valores nulos en rate_14_day
```

	country	country_code	continent	population	indicator	weekly_count	year_week	rate_14_day	cumulative_count	source
22778	Malta	MLT	Europe	514564	deaths	0	2020-01	NaN	0	TESSy COVID-19
9606	Czechia	CZE	Europe	10693939	cases	0	2020-01	NaN	0	TESSy COVID-19
28826	Portugal	PRT	Europe	10295909	cases	0	2020-01	NaN	0	TESSy COVID-19
25444	Netherlands	NLD	Europe	17407585	deaths	0	2020-01	NaN	0	TESSy COVID-19
3920	Belgium	BEL	Europe	11522440	deaths	0	2020-01	NaN	0	TESSy COVID-19
5900	Bulgaria	BGR	Europe	6951482	cases	0	2020-01	NaN	0	TESSy COVID-19
5996	Bulgaria	BGR	Europe	6951482	deaths	0	2020-01	NaN	0	TESSy COVID-19
28922	Portugal	PRT	Europe	10295909	deaths	0	2020-01	NaN	0	TESSy COVID-19
16806	Hungary	HUN	Europe	9769526	deaths	0	2020-01	NaN	0	TESSy COVID-19
21620	Luxembourg	LUX	Europe	602109	cases	0	2020-01	NaN	0	TESSy COVID-19

Figura 20. Casos Covid-19 EU countries

Los NaN en *rate_14_day* coinciden con la primera semana de 2020 y *weekly_count* igual a cero. Fecha en la que aún no se registraban casos de Covid-19 en Europa por lo que es apropiado reemplazar estos valores nulos con 0

- Elimino columnas:

Con el DF libre de valores nulos y con los datos solo de los países de la UE, elimino la variable *continent* pues sobra decir que para cada entrada el continente es Europa, también elimino la variable *source*, que para este caso en particular resulta indiferente

- Convertir variable *year_week* a datetime:

La variable *year_week* es de tipo object, la convierto en datetime con *pd.to_datetime*, obtengo una nueva variable con formato Año-Mes-Día, mostrando el último día de la semana reportada

- DF Limpio

```
DF limpio
Nº de filas / entradas: 5184
Nº de columnas / variables: 11
Tipo de datos para cada variable:
country                      object
country_code                  object
population                   int64
indicator                     object
weekly_count                  int64
year_week                     object
rate_14_day                   float64
cumulative_count              int64
Year                          object
week                          object
End_week                      datetime64[ns]
dtype: object
Total valores nulos: 0
```

II. Datos de vacunación de Covid-19 en la Unión Europea y el Espacio Económico Europeo

- ✓ *Data on COVID-19 vaccination in the EU/EEA*

- *Filtro:*

1. DF con datos de países de UE y EEE, así como regiones de algunos países, filtro los datos de los 27 países miembros de la UE.
2. Para el objeto de este estudio no me hacen falta los datos sobre vacunación por rangos de edad, por lo que filtro en la variable TargetGroup por 'ALL' población mayor de 18 años

```

# filtrar por los países miembros de la UE
from MyFunctions import EUcountries

Vaccine = Vaccine[(Vaccine.ReportingCountry.isin(EUcountries().Code))\
& (Vaccine.Region.isin(EUcountries().Code))]

# Filter target group 'ALL' >18 years
Vaccine = Vaccine[(Vaccine.TargetGroup == "ALL")]

#Verifco que estén los 27
print(Vaccine['ReportingCountry'].unique())
print(Vaccine['Region'].unique())

['AT' 'BE' 'BG' 'CY' 'CZ' 'DE' 'DK' 'EE' 'EL' 'ES' 'FI' 'FR' 'HR' 'HU'
 'IE' 'IT' 'LT' 'LU' 'LV' 'MT' 'NL' 'PL' 'PT' 'RO' 'SE' 'SI' 'SK']
['AT' 'BE' 'BG' 'CY' 'CZ' 'DE' 'DK' 'EE' 'EL' 'ES' 'FI' 'FR' 'HR' 'HU'
 'IE' 'IT' 'LT' 'LU' 'LV' 'MT' 'NL' 'PL' 'PT' 'RO' 'SE' 'SI' 'SK']

```

Figura 21. Filtro Vaccination +18

- Tratamiento valores nulos:

Después de filtrar quedan 4420 valores nulos en la variable *FirstDoseRefused* y 563 valores nulos en la variable *NumberDosesReceived*

Los valores que no son nulos en la variable *FirstDoseRefused* son 325 de 4745 registros, el 97% son valores nulos y al no ser información relevante puedo eliminar la columna.

Para el caso de la variable *NumberDosesReceived* son menos los valores nulos, pero al no ser datos relevantes para este estudio elimino también la columna

- Elimino columnas:

Elimino columnas que no necesito:

- *FirstDoseRefused* y *NumberDosesReceived* por lo expuesto anteriormente
- *Region* porque después de filtrar son los mismos valores de *ReportingCountry*
- *TargetGroup* porque después de filtrar solo tengo población mayor de 18 años

- Convertir variable *YearWeekISO* a datetime:

La variable *YearWeekISO* es de tipo object, la convierto en datetime con *pd.to_datetime*, obtengo una nueva variable con formato Año-Mes-Día, mostrando el último día de la semana reportada

- DF Limpio

```

DF limpio
Nº de filas / entradas: 4745
Nº de columnas / variables: 11
Tipo de datos para cada variable:
YearWeekISO           object
FirstDose              int64
SecondDose             int64
UnknownDose            int64
Population             int64
ReportingCountry       object
Vaccine                object
Denominator            int32
Year                   object
week                  object
End_week               datetime64[ns]
dtype: object

```

III. Datos relativos al valor agregado bruto por industria

✓ *Gross value added and income A*10 industry breakdowns*

- *Filtro:*

1. DF con datos desde el primer trimestre del año 1.975, para este DF filtro desde el primer trimestre del año 2007 por ser el último año en el que algún país ingresó formalmente a la UE y hasta el segundo semestre de 2021, ya que a la fecha se tienen demasiado valores nulos para el tercer trimestre de 2021
2. Filtro los datos de los 27 países miembros de la UE.
3. Por valores en cada variable: *From pandas to Pyspark*

```

#De este DataFrame quiero para cada variable
unitconcept = ['PYP_MNAC', 'PYP_MEUR'] #Todos excepto Previous year prices, million units of national currency (PYP_MNAC)
                                         # y Previous year prices, million euro (PYP_MEUR)
s_adjconcept= ['SCA'] #Seasonally and calendar adjusted data
na_itemconcept= ['B1G'] #Value added, gross
gdp_Ind_count = gdp_Ind.filter((gdp_Ind.s_adj.isin(s_adjconcept)) \
                                & (gdp_Ind.na_item.isin(na_itemconcept))\
                                & (~gdp_Ind.unit.isin(unitconcept)))

```

Figura 21. Filtro valor agregado bruto por industria

- Elimino columnas:

- s_adj, porque después de filtrar como ajuste estacional tomo ‘SCA’ datos ajustados a temporada y calendario
- na_item, porque después de filtrar como indicador de cuentas nacionales tomo ‘B1G’ valor añadido bruto

- Unpivot DF:

Traspaso todas las columnas relativas a fechas en una sola variable, Date

- DF Limpio

Resumen:

```
gdp_Ind_F.summary().show()
```

summary	unit	nace_r2	country	Date	cant	EconomicAct	CountryName
count	505992	505992	505992	505992	505992	505992	505992
mean	null	null	null	null	14514.765444943794	null	null
stddev	null	null	null	null	164987.31691812963	null	null
min	CLV05_MEUR	A	AT	2007Q1	-60.9	Agriculture, fore...	Austria
25%	null	null	null	null	3.0	null	null
50%	null	null	null	null	101.056	null	null
75%	null	null	null	null	483.0	null	null
max	PD_PCH_SM_NAC	TOTAL	SK	2021Q2	1.1509848E7	Wholesale and ret...	Sweden

Figura 22. DF PIB por industria limpia

Valores nulos: 0

unit	nace_r2	country	Date	cant	EconomicAct	CountryName
0	0	0	0	0	0	0

IV. Datos relativos a población y empleo

✓ Population and employment

- Filtro:

1. DF con datos desde el primer trimestre del año 1.975, para este DF filtro desde el primer trimestre del año 2018 porque quiero determinar como el Covid-19 ha afectado el empleo, teniendo en cuenta que los primeros casos en Europa se detectaron durante los primeros meses del año 2020 dejó los períodos de 2018 y 2019 como punto de comparación; y hasta el segundo semestre de 2021, ya que a la fecha se tienen demasiado valores nulos para el tercer trimestre de 2021
2. Filtro los datos de los 27 países miembros de la UE.
3. Por valores en cada variable: *From pandas to Pyspark*

```
#De este DataFrame quiero todos los datos con NSA y THS_PER
#Elimino la columna s_adj y unit

s_adjconcept= ['NSA'] #Unadjusted data (i.e. neither seasonally adjusted nor calendar adjusted data)
unitconcept = ['THS_PER'] # Thousand persons

PopAndEmp = PopAndEmp.filter((PopAndEmp.s_adj.isin(s_adjconcept)) & (PopAndEmp.unit.isin(unitconcept)))
PopAndEmp = PopAndEmp.select([c for c in PopAndEmp.columns if c not in {'s_adj'}])
PopAndEmp = PopAndEmp.select([c for c in PopAndEmp.columns if c not in {'unit'}])
```

Figura 23. Filtro empleo y población

- Elimino columnas:
 - s_adj, porque después de filtrar como ajuste estacional tomo ‘NSA’ datos sin ajustar
 - unit, porque después de filtrar como unidad de medida tomo ‘THS_PER’ miles de personas

- Unpivot DF:

Traspaso todas las columnas relativas a fechas en una sola variable, Date

- DF Limpio

Resumen:

```
PopAndEmp.summary().show()
```

	summary	na_item	country	Date	cant
count	2646	2646	2646		2646
mean	null	null	null	6750.356636432363	
stddev	null	null	null	12111.722871965178	
min	EMP_DC	AT	2018Q1		21.39
25%	null	null	null		655.07
50%	null	null	null		2633.2
75%	null	null	null		5458.74
max	SELF_NC	SK	2021Q2		83194.0

Figura 24. DF empleo y población limpia

Valores nulos: 0

	na_item	country	Date	cant
	0	0	0	0

V. Datos relativos a población por sexo, edad, ciudadanía y empleo

- ✓ *Population by sex, age, citizenship and labour status*

- Filtro:

1. DF con datos desde el primer trimestre del año 1.998 hasta el segundo trimestre del año 2021, para este DF filtro desde el primer trimestre del año 2018 porque quiero determinar como el Covid-19 ha afectado el empleo, teniendo en cuenta que los primeros casos en

Europa se detectaron durante los primeros meses del año 2020 dejó los períodos de 2018 y 2019 como punto de comparación;

2. Filtro los datos de los 27 países miembros de la UE.
3. Por valores en cada variable: *From pandas to Pyspark*

```
#De este DataFrame elimino la columna unit porque solo tiene un valor: thousands
#La variable age la dejo tal cual
#De las variables citizen solo me interesan TOTAL, filtro por esta y elimino la columna
#De la variable age me interesa la población en edad de trabajar, entre 15 y 64, y por rangos de edades
#De la variable wstatus todos menos unknown

citizenconcept = ['TOTAL']
ageconcept = ['Y15-64', 'Y20-64', 'Y20-24', 'Y25-29', 'Y30-34', 'Y35-39', 'Y40-59']
wstatusconcept = ['POP', 'INAC', 'ACT', 'EMP', 'UNE']

Pop_Age = Pop_Age.filter((Pop_Age.citizen.isin(citizenconcept)) & (Pop_Age.age.isin(ageconcept)) & (Pop_Age.wstatus.isin(wstatusconcept)))
Pop_Age = Pop_Age.select([c for c in Pop_Age.columns if c not in {'unit', 'citizen'}])
```

Figura 25. Filtro DF población por sexo, edad, ciudadanía y empleo

- Elimino columnas:
 - unit porque solo tiene una unidad de medida ‘THS’, Thousand
 - citizen, porque después de filtrar, como nacionalidad tomo ‘TOTAL’, para este caso me es indiferente
- Unpivot DF:

Traspaso todas las columnas relativas a fechas en una sola variable, Date

- DF Limpio

Resumen:

```
Pop_Age.summary().show()
```

summary	sex	age	wstatus	country	Date	cant
count	39690	39690	39690	39690	39690	39690
mean	null	null	null	null	null	NaN
stddev	null	null	null	null	null	NaN
min	F	Y15-64	ACT	AT	2018Q1	0.9
25%	null	null	null	null	null	63.4
50%	null	null	null	null	null	276.4
75%	null	null	null	null	null	1287.6
max	T	Y40-59	UNE	SK	2021Q2	NaN

Figura 26. DF población por sexo, edad, ciudadanía y empleo limpio

Valores nulos:

Varios, Alemania no ha reportado datos para el año 2021. Lo trato en la siguiente fase

sex	age	wstatus	country	2021Q2	2021Q1	2020Q4	2020Q3	2020Q2	2020Q1	2019Q4	2019Q3	2019Q2	
	2018Q1												
T Y15-64	ACT	DE	41581.0 41789.7	NaN	NaN	NaN	NaN	42938.5 42498.6	42002.6				
42019.5													
T Y15-64	EMP	DE	40059.4 40036.8	NaN	NaN	NaN	NaN	41599.4 41153.3	40681.2				
40466.3													
T Y15-64	INAC	DE	11429.3 11680.0	NaN	NaN	NaN	NaN	11053.3 10952.0	11200.8				
11627.7													
T Y15-64	POP	DE	53010.4 53469.6	NaN	NaN	NaN	NaN	53991.9 53450.6	53203.4				
53647.2													
T Y15-64	UNE	DE	1521.6 1752.9	NaN	NaN	NaN	NaN	1339.1 1345.3	1321.4				
1553.2													
T Y20-24	ACT	DE	3261.5 3303.6	NaN	NaN	NaN	NaN	3171.4 3116.0	3131.4				
3041.4													
T Y20-24	EMP	DE	3058.0 3031.9	NaN	NaN	NaN	NaN	3012.3 2939.4	2976.3				
2851.5													
T Y20-24	INAC	DE	1368.5 1380.2	NaN	NaN	NaN	NaN	1299.7 1230.7	1304.8				

Figura 27. Valores nulos Alemania DF población por sexo, edad, ciudadanía y empleo limpio

VI. Datos relativos a empleo por actividad económica

✓ Employment A*10 industry breakdowns

- Filtro:

1. DF con datos desde el primer trimestre del año 1.975 hasta el segundo trimestre del año 2021, para este DF filtro desde el primer trimestre del año 2018 porque quiero determinar como el Covid-19 ha afectado el empleo, teniendo en cuenta que los primeros casos en Europa se detectaron durante los primeros meses del año 2020 dejó los periodos de 2018 y 2019 como punto de comparación y hasta el segundo trimestre del año 2021 porque para el tercer trimestre a la fecha se registran demasiados valores nulos
2. Filtro los datos de los 27 países miembros de la UE.
3. Por valores en cada variable: *From pandas to Pyspark*

```
# Para la variable unit me quedo con thousands
# s_adj solo not adjusted data
# De la variable na_item me quedo con total employment domestic concept
# Elimino las tres variables filtradas

unitconcept = ['THS_PER'] #Thousand persons
s_adjconcept= ['NSA'] #not adjusted data
na_itemconcept= ['EMP_DC'] #Total employment domestic concept
empByInd = empByInd.filter((empByInd.unit.isin(unitconcept)) & (empByInd.s_adj.isin(s_adjconcept)) & (empByInd.na_item.isin(na_itemconcept)))
empByInd = empByInd.select([c for c in empByInd.columns if c not in {'unit', 's_adj', 'na_item'}])
```

Figura 28. Filtro DF empleo por actividad económica

- Elimino columnas:

- unit porque después de filtrar como unidad de medida tomo ‘THS_PER’ miles de personas
- s_adj, porque después de filtrar como ajuste estacional tomo ‘NSA’ datos sin ajustar
- na_item, porque después de filtrar como indicador de cuentas nacionales tomo ‘EMP_DC’ total empleo concepto interno

- Unpivot DF:

Traspaso todas las columnas relativas a fechas en una sola variable, Date

- DF Limpio

Resumen:

```
empByInd.summary().show()
```

summary	nace_r2	country	Date	cant
count	4536	4536	4536	4536
mean	null	null	null	1372.0216578483232
stddev	null	null	null	3843.64418314649
min	A	AT	2018Q1	2.09
25%	null	null	null	83.2
50%	null	null	null	286.23
75%	null	null	null	998.3
max	TOTAL	SK	2021Q2	45559.0

Figura 29. DF empleo por actividad económica limpio

Valores nulos: 0

nace_r2	country	Date	cant
0	0	0	0

5.5. Análisis de variables

I. Datos número de casos y muertes por COVID-19



Figura 30. Correlación de datos casos Covid-19

En el mapa de calor se puede detectar que las variables población y ratio de 14 días no tienen relación significativa, su valor se acerca a 0; ninguna de las variables tiene una fuerte relación lineal positiva; la relación más intensa en este conjunto de datos se las variables *cumulative_count* y *weekly_count*, lo visualizo en un gráfico de dispersión:

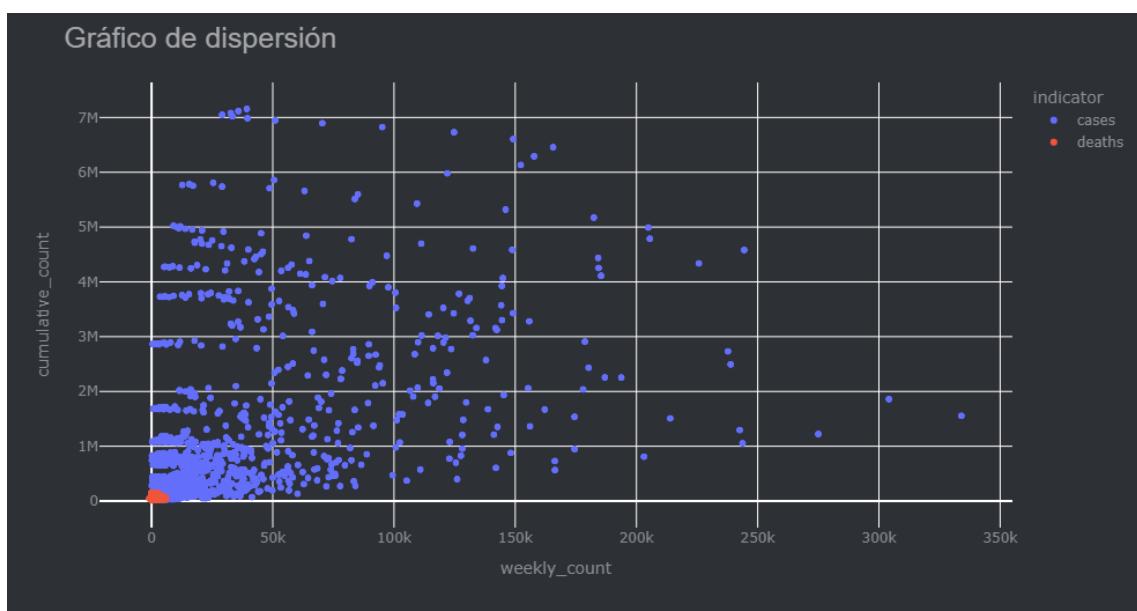


Figura 31. Gráfico de dispersión casos Covid-19

En el siguiente grafico de cajas se ve a primera vista la relación de datos entre las variables ratio 14 días e indicador.

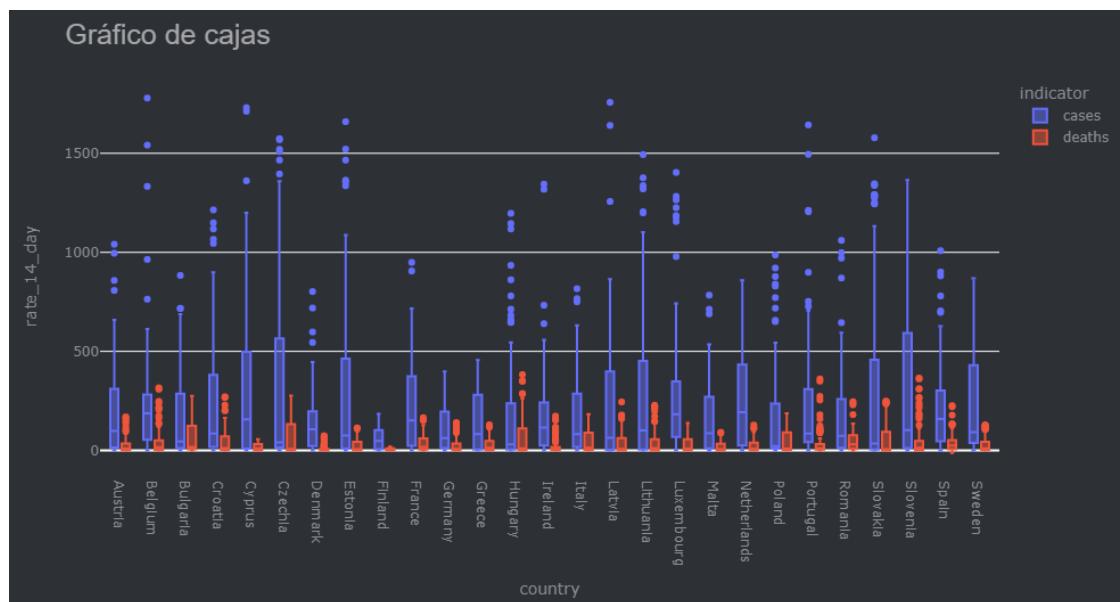


Figura 32. Gráfico de cajas casos Covid-19

La media de casos y muertes en los 27 países es cercana, sin embargo países como Bélgica, Letonia y Portugal presentan outliers altos en relación con países como Finlandia, Alemania y Grecia en los que no se tiene presencia de outliers.

II. Datos de vacunación de Covid-19 en la Unión Europea y el Espacio Económico Europeo

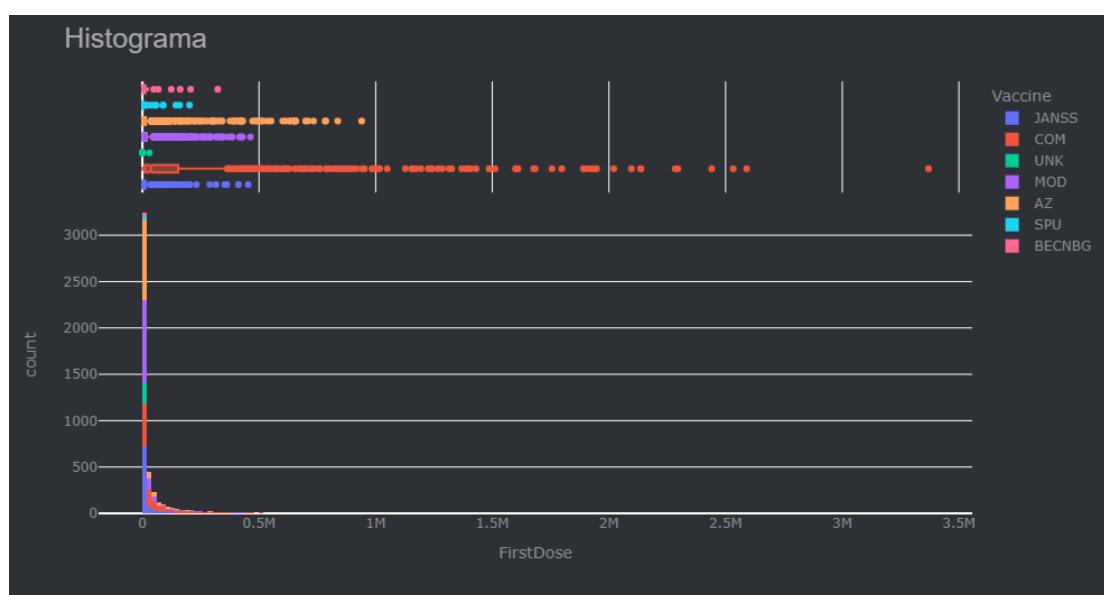


Figura 33. Histograma vacunación

III. Datos relativos al valor agregado bruto por industria

✓ Gross value added and income A*10 industry breakdowns

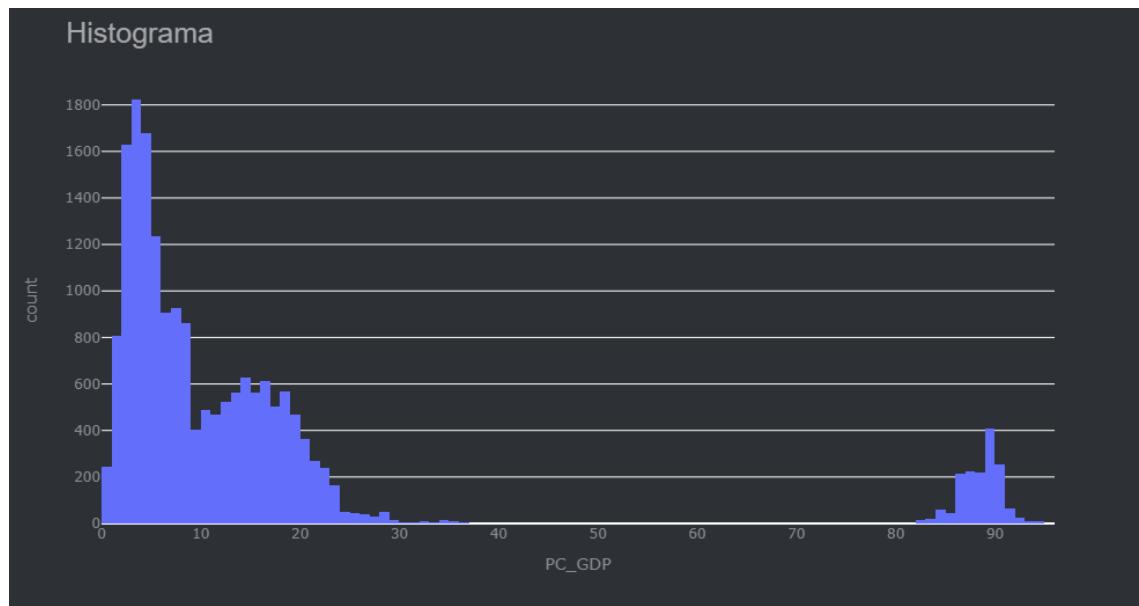


Figura 34. Histograma PIB por industria

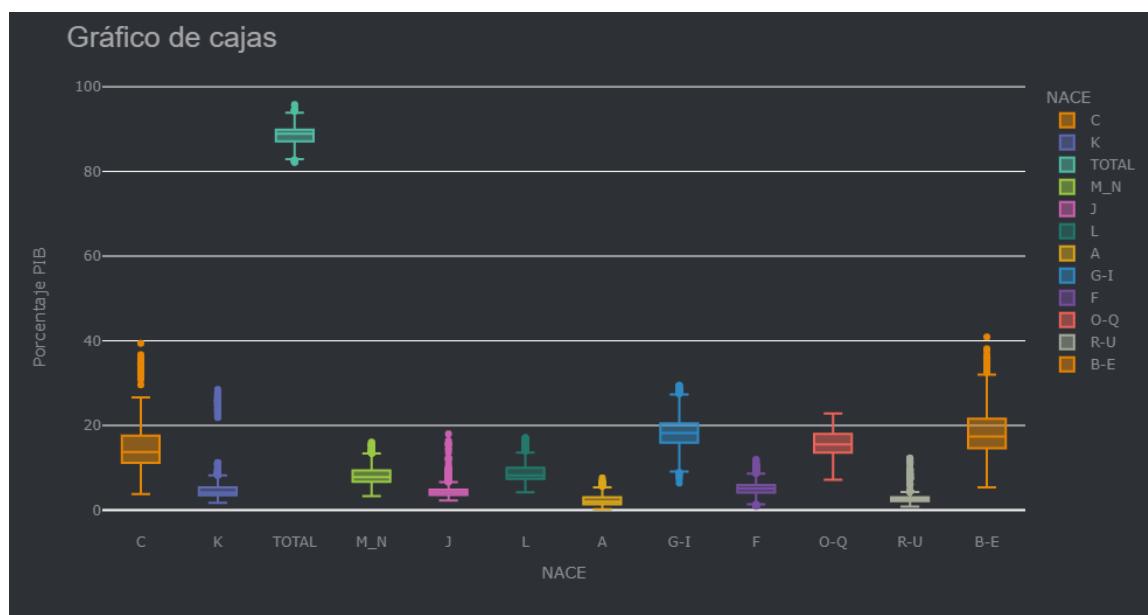


Figura 35. Gráfico de cajas PIB por industria

5.6. Transformación de datos

Una vez que se ha dado fin al proceso de la limpieza y preparación de los DFs y estos han sido almacenados en HDFS, se prosigue con la fase de transformación de datos, a continuación se especifican los datos transformados, DFs de origen y proceso de transformación

» Conversión datos semanales a datos trimestrales

$$\text{YearWeek} = \text{YearQuarter}$$

De los Data Frames seleccionados para este trabajo hay dos que vienen con formato fecha YearWeek y cuatro con formato fecha YearQuarter; para equiparar datos convierto los DF que tienen datos semanales a datos trimestrales

+ DFs:

- I. Datos número de casos y muertes por COVID-19, limpio: Weeklycases
- II. Datos de vacunación de Covid-19 en la Unión Europea y el Espacio Económico Europeo, limpio: WeeklyVaccineData

+ Proceso:

- Leer datos desde ruta HDFS

```
cases14day = spark.read.parquet("hdfs://localhost:9000/TFM_CEE/raw/Weeklycases.parquet")
```

```
VaccinationQ = spark.read.parquet("hdfs://localhost:9000/TFM_CEE/raw/WeeklyVaccineData.parquet")
```

- Extraer quarter y año from date

Con la función *quarter* extraigo el trimestre de la columna date, en una nueva variable '*quarter*'

Con la función *year* extraigo el año de la columna date, en una nueva variable '*Real_year*', no utilizo la columna '*year*' que ya tengo porque la última semana de 2020 es en realidad la primera de 2021 en modo ISO

```
casesQ = cases14day.withColumn('quarter', func.quarter(cases14day.date))
casesQ = casesQ.withColumn('Real_year', func.year(cases14day.date))
```

```
VaccinationQ = Vaccination.withColumn('quarter',
func.quarter(Vaccination.date))
```

```
VaccinationQ = VaccinationQ.withColumn('Real_year',
func.year(VaccinationQ.date))
```

- Concatenar en una columna:

Con la fusión concat_ws junto los valores de ‘quarter’ y ‘Real_Year’ en una nueva variable ‘DateQ’

```
casesQ.withColumn('DateQ', func.concat_ws("Q", casesQ.Real_year, casesQ.quarter))
```

quarter	Real_year	DateQ
1	2020	2020Q1
1	2020	2020Q1

- Agrupar datos por trimestre

Una vez asignado el trimestre a cada fecha agrupo los datos por DateQ y obtengo:

Para datos Covid:

En weekly_count suma de casos dentro del trimestre

En cumulative_count el dato más alto en determinado trimestre

En Para rate_14_day la media para cada trimestre

```
CovidCasesQ = casesQ.groupBy('country_code', 'country', 'DateQ',
                             'population', 'indicator') \
    .agg(sum('weekly_count').alias('Quarter_count'), \
         max('cumulative_count').alias('cumulative_count'), \
         avg('rate_14_day').alias('rate_Quarter')) \
    .sort('country', 'indicator', 'DateQ')
```

Selecciono muestra de datos de España:

```
CovidCasesQ.filter(CovidCasesQ.country == 'Spain').show(3)
```

country_code	country	DateQ	population	indicator	Quarter_count	cumulative_count	rate_Quarter
ESP	Spain	2020Q1	47332614	cases	157679	157679	44.09124427723027
ESP	Spain	2020Q2	47332614	cases	100538	258217	39.433535356663576
ESP	Spain	2020Q3	47332614	cases	543499	801716	165.47884200619376

Figura 36. Muestra conversión año a trimestre DF Covid-19

Para datos Vacunación:

Para cada variable, con suma el total vacunas aplicadas por trimestre

```
VaccinationQ = VaccinationQ.groupBy('ReportingCountry', 'DateQ',
                                         'Vaccine', 'Denominator') \
    .agg(sum('FirstDose').alias('FirstDose'), \
         sum('SecondDose').alias('SecondDose'), \
         sum('UnknownDose').alias('UnknownDose')) \
    .sort('ReportingCountry', 'Vaccine', 'DateQ')
```

Selecciono muestra de datos de Italia:

```
#DF Vaccination por trimestre
VaccinationQ.filter(VaccinationQ.ReportingCountry == 'IT').show(3)
```

ReportingCountry	DateQ	Vaccine	Denominator	FirstDose	SecondDose	UnknownDose
IT	2021Q1	AZ	50208329	2206753	973	0
IT	2021Q2	AZ	50208329	4195214	2719313	0
IT	2021Q3	AZ	50208329	27517	2817909	0

only showing top 3 rows

Figura 36. Muestra conversión año a trimestre DF Vacunación

» Porcentaje de casos positivos y muertes a causa del Covid-19

+ DFs:

* Datos número de casos y muertes por COVID-19, limpio: Weeklycases

$$X = \frac{\text{Quarter_count} * 100}{\text{Population}}$$

Quarter_count = Número de personas casos positivo o muerte por trimestre

Population = Total de población para cada trimestre

+ Proceso:

```
CovidPercQ =
CovidCasesQ.withColumn('cumulative_perc',
                         (CovidCasesQ.cumulative_count * 100) / CovidCasesQ.population) \
    .withColumn('Quarter_perc',
                (CovidCasesQ.Quarter_count * 100) / CovidCasesQ.population)
```

» Porcentaje de población

+ DF:

**Datos de vacunación de Covid-19 en la Unión Europea y el Espacio Económico Europeo, limpio: WeeklyVaccineData*

$$X = \frac{Full_vaccine * 100}{Denominator}$$

+ Proceso:

- Ratio de población mayor de 18 años con pauta completa de vacunación.

Calcular datos para cantidad de dosis completa vacunación

Para Janssen:

$$PautaC = FirstDose + UnknownDose$$

Para No Janssen:

$$PautaC = SecondDose$$

```
Fully_vaccine = when(col("Vaccine") == 'JANSS', \
    (VaccinationQ.FirstDose + VaccinationQ.UnknownDose)) \
    .when(col("Vaccine") != 'JANSS', VaccinationQ.SecondDose)

VaccinationQF = VaccinationQ.withColumn("Fully_vaccine", Fully_vaccine)
```

- Número de personas con pauta completa de vacunación para cada trimestre, tipo de vacuna indiferente

$$T = DateQ \left(\sum \text{Fully_vaccine} \right)$$

```
VaccinationQF = VaccinationQF.withColumnRenamed('ReportingCountry', \
    'CountryCode') \
    .groupBy('CountryCode', 'DateQ', 'Denominator') \
    .agg(sum('Fully_vaccine').alias('Full_vaccine')) \
    .sort('CountryCode', 'DateQ')
```

- Acumulado de personas con pauta completa de vacunación

```

windowval = (Window.partitionBy('CountryCode').
              .orderBy('CountryCode', 'DateQ')
              .rangeBetween(Window.unboundedPreceding, 0))
VaccineFull = VaccinationQF.withColumn('Full_vac_acumulado', func.sum('Full_vaccine').over(windowval))

```

- Porcentaje de población vacunada

```

VaccineFull = VaccineFull.withColumn('Vac_perc_acum', \
(VaccineFull.Full_vac_acumulado * 100) / VaccineFull.Denominator) \
.withColumn('Vac_perc', \
(VaccineFull.Full_vaccine * 100) / VaccineFull.Denominator)

```

» Top 3 actividades económicas con más aportación al PIB para cada país

+ DF:

* Datos relativos al valor agregado bruto por industria, limpio: gdpIndustries

+ Proceso:

- Leer datos desde ruta HDFS

```

gdp_Ind = spark.read.parquet("hdfs://localhost:9000/TFM_CEE/row/gdpIndustries.parquet")

```

- Filtros:

* En Variable unit 'PC_GDP' porcentaje de producto interno bruto

* En Variable nace_r2 'TOTAL' todas las actividades NACE menos el total de actividades

- Promedio:

El DF contiene datos desde el primer trimestre del año 2007, hallo el promedio de porcentaje de aportación al PIB por país, para cada actividad en el periodo de tiempo de 2007 hasta 2021, en nueva variable 'Prom_Historico'

- Ranking:

Para determinar cuales han sido las tres actividades económicas que mas aportan al PIB por país utilizo las funciones window y rank extraigo el top 3

```

# Visualizar top 3 Industrias porcentaje x pais
# Filtro en Unit: ['PC_GDP'] #Percentage of gross domestic product (GDP)
# Filtro en actividades Nace que no sean el total
# Avg cantidad = Promedio historico

Top3 = gdp_Ind.filter((gdp_Ind.unit == 'PC_GDP') & (gdp_Ind.nace_r2 != 'TOTAL')) \
    .select('CountryName', 'EconomicAct', 'nace_r2', 'Date', 'cant') \
    .groupBy('CountryName', 'nace_r2', 'EconomicAct') \
    .agg(func.avg('cant').alias('Prom_Historico')) \
    .orderBy('CountryName', 'Prom_Historico')

from pyspark.sql.window import Window
from pyspark.sql.functions import rank
window1 = Window.partitionBy(Top3['CountryName']).orderBy(Top3['Prom_Historico'].desc())

Top3PIB = Top3.select('*', rank().over(window1).alias('Top3')) \
    .filter(func.col('Top3') <= 3)
Top3PIB.show(8, truncate =False)

```

Figura 37. Código ranking aportación al PIB por actividad

Selecciono muestra de datos para España:

```
Top3PIB.filter(Top3PIB.CountryName == 'Spain').show()
```

CountryName	nace_r2	EconomicAct	Prom_Historico	Top3
Spain	G-I	Wholesale and ret...	20.67	1
Spain	O-Q	Public administrat...	16.81	2
Spain	B-E	Industry (except ...)	14.96	3

Figura 38. Muestra DF ranking aportación al PIB por actividad

» Estadísticas mercado laboral por país

+ DFs:

- * Datos relativos a población y empleo, limpio: PopAndEmp
- * Datos relativos a población por sexo, edad, ciudadanía y empleo, limpio: PopWStatusByAge
- * Datos relativos a empleo por actividad económica, limpio: EmpByIndust

+ Cargo datos:

```

popAndEmpl = spark.read.parquet("hdfs://localhost:9000//TFM_CEE/row/PopAn
dEmp.parquet")

popByAge = spark.read.parquet("hdfs://localhost:9000//TFM_CEE/row/PopWSta
tusByAge.parquet")

EmpByInd = spark.read.parquet("hdfs://localhost:9000//TFM_CEE/row/EmpByIn
dust.parquet")

```

- + Para hallar las tasas de actividad, ocupación y paro a partir del DF PopWStatusByAge creo un DF para cada sexo en edad de trabajar Y15-64 y lo pivoteo por la variable wstatus, situación laboral, obtengo DF sin distinción de sexo, otro DF con datos para hombres y otro con datos para mujeres.

PivotFemales_Y15to64.filter(PivotFemales_Y15to64.country == 'PT').show								
country	CountryName	Date	ACT	EMP	INAC	POP	UNE	
PT	Portugal	2018Q1	2472	2266	957	3429	206	
PT	Portugal	2018Q2	2479	2299	946	3425	180	
PT	Portugal	2018Q3	2489	2304	932	3421	185	
PT	Portugal	2018Q4	2469	2284	948	3416	185	
PT	Portugal	2019Q1	2481	2288	943	3424	193	
PT	Portugal	2019Q2	2494	2324	927	3421	170	
PT	Portugal	2019Q3	2488	2312	929	3417	176	
PT	Portugal	2019Q4	2507	2315	908	3416	192	
PT	Portugal	2020Q1	2484	2299	943	3426	185	
PT	Portugal	2020Q2	2380	2243	1046	3425	137	
PT	Portugal	2020Q3	2488	2285	936	3424	203	
PT	Portugal	2020Q4	2489	2299	938	3427	190	
PT	Portugal	2021Q1	2435	2252	976	3411	183	
PT	Portugal	2021Q2	2486	2314	923	3409	172	

Figura 39. Muestra DF situación laboral mujeres para Portugal

» Tasa de Actividad

$$TA = \left(\frac{ACT}{POP} \right) * 100$$

ACT = Cantidad población activa

POP = Total población en edad de trabajar, de 15 a 64 años

- Aplico la fórmula de TA para cada uno de los tres DFs pivoteados

```

# TASA DE ACTIVIDAD
# indice que mide el nivel de actividad en el empleo de un país
# Calcular la tasa de actividad, para el total de la población, hombres y mujeres.
# Se calcula como el cociente entre la población activa (ACT) y la población en edad de trabajar (POP) y15-64.
# Tasa de empleo = (ACT/POP)*100

TAall = PivotAll_Y15to64.withColumn('Total', (PivotAll_Y15to64.ACT / PivotAll_Y15to64.POP)*100 ) \
    .select('country', 'CountryName', 'Date', 'Total')

TAmales = PivotMales_Y15to64.withColumn('Hombres', (PivotMales_Y15to64.ACT / PivotMales_Y15to64.POP)*100 ) \
    .select('country', 'CountryName', 'Date', 'Hombres')
|
TAfemales = PivotFemales_Y15to64.withColumn('Mujeres', (PivotFemales_Y15to64.ACT / PivotFemales_Y15to64.POP)*100 ) \
    .select('country', 'CountryName', 'Date', 'Mujeres')

```

Figura 40. Código Formula Tasa Actividad

- Con estos resultados hago un join con los tres DFs
- Obtengo un DF final con la TA general y con distinción por hombres y mujeres

» Tasa de Paro o desempleo

$$TP = \left(\frac{UNE}{ACT} \right) * 100$$

UNE = Cantidad población desempleada,

ACT = Cantidad población activa

- Aplico la fórmula de TO para cada uno de los tres DFs pivoteados
- Con estos resultados hago un join con los tres DFs

```

#create Final DF Tasa de Paro Joining the 3 dataframe

TP_F = TPall.join(TPmales, on=['country', 'CountryName', 'Date'], how='outer')\
    .join(TPfemales, on=['country', 'CountryName', 'Date'], how='outer')\
    .orderBy('country', 'Date')

```

Figura 40. Join en DFs Tasa de Paro

- Obtengo un DF final con la TO general y con distinción por hombres y mujeres

» Tasa de Ocupación o empleo

$$TO = \left(\frac{EMP}{POP} \right) * 100$$

EMP = Cantidad población empleada,

POP = Total población en edad de trabajar, de 15 a 64 años

- Aplico la fórmula de TO para cada uno de los tres DFs pivoteados

- Con estos resultados hago un join con los tres DFs
- Obtengo un DF final con la TO general y con distinción por hombres y mujeres

```
TO_F.show()
```

country	CountryName	Date	Total	Hombres	Mujeres
AT	Austria	2018Q1	71.9627778735137	75.77639751552795	68.15834767641996
AT	Austria	2018Q2	73.00275482093664	77.67057201929704	68.34136269786649
AT	Austria	2018Q3	73.8209982788296	78.56159669649003	69.05335628227195
AT	Austria	2018Q4	73.27141382868938	77.78925619834712	68.76288659793815
AT	Austria	2019Q1	72.72883688919477	76.54150878401653	68.92402887590238
AT	Austria	2019Q2	73.39402267262109	77.98007557540365	68.77361731363793
AT	Austria	2019Q3	74.24737656975744	78.91836031691354	69.58762886597938
AT	Austria	2019Q4	73.84219554030875	78.37281153450051	69.31779225231402
AT	Austria	2020Q1	72.32326698695951	76.15939539677086	68.4950291395269
AT	Austria	2020Q2	71.05850060044605	75.10302197802197	66.99794379712132

Figura 41. Muestra DF Tasa de Ocupación

» Tasa de Ocupación por Actividad económica

$$TA = \left(\frac{EMP_NACE}{POP} \right) * 100$$

EMP_NACE = Cantidad población empleada en cada actividad económica NACE

POP = Total población en edad de trabajar, de 15 a 64 años

- Selecciono el DF EmpByInd y lo pivoteo por la variable nace_r2, así obtengo datos de empleo por miles para cada actividad económica
- Hago un join con el DF creado anteriormente PivotAll_Y15to64 ya que contiene datos de población en edad de trabajar para cada trimestre y país
- Obtengo un DF llamado TObyI y aplico la formula TA sobre cada una de las actividades NACE

```

T0xInd = T0byI.withColumn('Tdo_A', (T0byI.A / T0byI.POP)*100 ) \
    .withColumn('Tdo_BE', (T0byI.BE / T0byI.POP)*100 ) \
    .withColumn('Tdo_C', (T0byI.C / T0byI.POP)*100 ) \
    .withColumn('Tdo_F', (T0byI.F / T0byI.POP)*100 ) \
    .withColumn('Tdo_GI', (T0byI.GI / T0byI.POP)*100 ) \
    .withColumn('Tdo_J', (T0byI.J / T0byI.POP)*100 ) \
    .withColumn('Tdo_K', (T0byI.K / T0byI.POP)*100 ) \
    .withColumn('Tdo_L', (T0byI.L / T0byI.POP)*100 ) \
    .withColumn('Tdo_MN', (T0byI.MN / T0byI.POP)*100 ) \
    .withColumn('Tdo_OQ', (T0byI.OQ / T0byI.POP)*100 ) \
    .withColumn('Tdo_RU', (T0byI.RU / T0byI.POP)*100 ) \
    .select('country', 'CountryName', 'Date', 'Tdo_A', 'Tdo_BE', \
        'Tdo_C', 'Tdo_F', 'Tdo_GI', 'Tdo_J', \
        'Tdo_K', 'Tdo_L', 'Tdo_MN', 'Tdo_OQ', 'Tdo_RU',)

```

Figura 42. Código Formula Tasa Ocupación actividades NACE

5.7. Almacenamiento

Los datos se guardan en HDFS en formato parquet; este “es el formato de almacenamiento en columnas principal en el ecosistema Hadoop. Fue desarrollado por primera vez por Twitter y Cloudera en cooperación” (apache.org, 2020). Es un formato de almacenamiento columnar que admite estructuras anidadas y se utiliza bastante en escenarios OLAP, online analytical processing.

Las ventajas de este formato se dan en el almacenamiento, ya que el almacenamiento de columnas facilita la comprensión de las mismas lo que reduce el espacio en disco; también tiene ventajas en rendimiento ya que permite leer solo las columnas requeridas y omitir aquellas que no cumplen condiciones.

Uso este formato en este caso porque además de sus ventajas es open source y uno de los formatos más utilizados en el big data

Para el almacenamiento de los datos he creado el directorio TFM_CEE Covid Economic Effects, en la siguiente ruta:

```

C:\hadoop\sbin>hdfs dfs -ls /TFM_CEE
Found 2 items
drwxr-xr-x  - icasa supergroup          0 2021-11-06 13:21 /TFM_CEE/output
drwxr-xr-x  - icasa supergroup          0 2021-11-06 13:21 /TFM_CEE/row

C:\hadoop\sbin>

```

Figura 43. Ruta directorio HDFS

Dentro de este he creado dos directorios ‘row’ y ‘output’; en el directorio ‘row’ almaceno los datos una vez limpios y preparados, pero sin transformaciones, como row data

Una vez cargados los datos originales desde el directorio row, los transformo obteniendo nuevos datos pertinentes para el cumplimiento de los objetivos de este trabajo en nuevos DFS, estos los almaceno en el subdirectorio output

```
C:\hadoop\sbin>hdfs dfs -ls /TFM_CEE/row
C:\hadoop\sbin>hdfs dfs -ls /TFM_CEE/output
```

Figura 44. Directorios HDFS vacíos

```
C:\hadoop\sbin>hdfs dfs -ls /TFM_CEE/row
Found 8 items
drwxr-xr-x  - icasa supergroup      0 2021-11-09 14:32 /TFM_CEE/row/EmpByIndust.parquet
drwxr-xr-x  - icasa supergroup      0 2021-11-08 02:52 /TFM_CEE/row/PopAndEmp.parquet
drwxr-xr-x  - icasa supergroup      0 2021-11-09 14:12 /TFM_CEE/row/PopWStatusByAge.parquet
drwxr-xr-x  - icasa supergroup      0 2021-11-06 21:03 /TFM_CEE/row/QuarterVaccineData.parquet
drwxr-xr-x  - icasa supergroup      0 2021-11-06 21:02 /TFM_CEE/row/WeeklyVaccineData.parquet
drwxr-xr-x  - icasa supergroup      0 2021-11-07 23:34 /TFM_CEE/row/Weeklycases.parquet
drwxr-xr-x  - icasa supergroup      0 2021-11-08 00:08 /TFM_CEE/row/gdpIndustries.parquet
drwxr-xr-x  - icasa supergroup      0 2021-11-07 23:34 /TFM_CEE/row/quartercases.parquet

C:\hadoop\sbin>
```

Figura 45. Directorio row con datos limpios

```
C:\hadoop\sbin>hdfs dfs -ls /TFM_CEE/output
Found 7 items
drwxr-xr-x  - icasa supergroup      0 2021-11-09 21:30 /TFM_CEE/output/TasaAct.parquet
drwxr-xr-x  - icasa supergroup      0 2021-11-09 21:32 /TFM_CEE/output/TasaOcup.parquet
drwxr-xr-x  - icasa supergroup      0 2021-11-09 21:40 /TFM_CEE/output/TasaOcupByInd.parquet
drwxr-xr-x  - icasa supergroup      0 2021-11-09 21:31 /TFM_CEE/output/TasaParo.parquet
drwxr-xr-x  - icasa supergroup      0 2021-11-10 00:19 /TFM_CEE/output/Top3PIB_Ind.parquet
drwxr-xr-x  - icasa supergroup      0 2021-11-07 23:34 /TFM_CEE/output/casesporcenpobl.parquet
drwxr-xr-x  - icasa supergroup      0 2021-11-06 23:34 /TFM_CEE/output/completeVaccine.parquet

C:\hadoop\sbin>
```

Figura 46. Directorio output con datos procesados

5.8. Visualización y resultados

En este apartado se presentan en formato imagen los gráficos de mayor relevancia, resultado de la transformación de datos del proceso anterior. En el archivo *Graphics* anexo a este documento se encuentran la totalidad de gráficos interactivos generados por Plotly

Covid-19

Para los datos relativos al Covid-19 se presentan los datos como porcentaje del total de la población mayor de 18 años, casos positivos, muertes y pauta completa de vacunación.

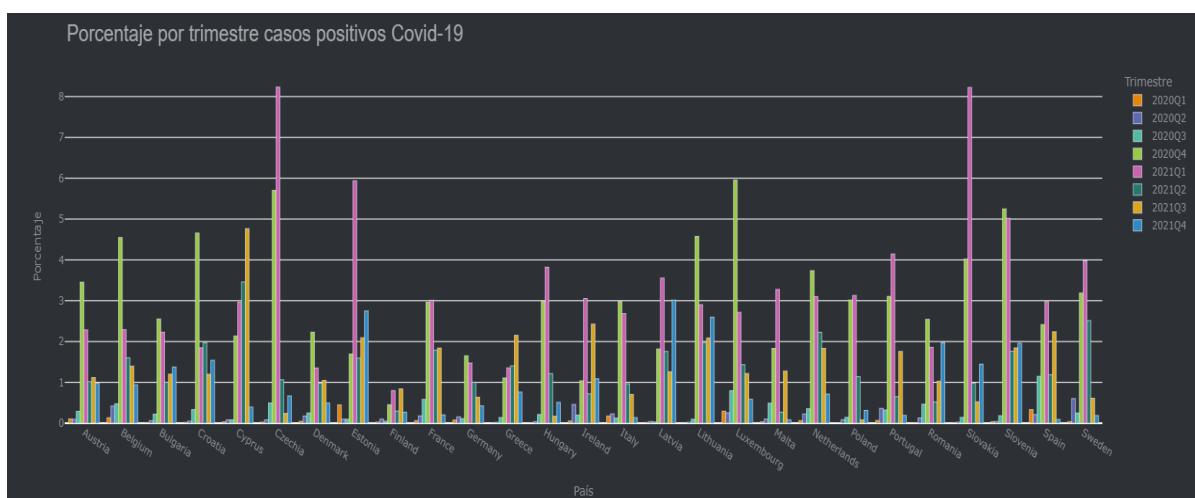


Figura 47. Porcentaje población casos positivos

El grafico anterior muestra el porcentaje de casos positivos por población en cada uno de los 27 países de la UE durante los ocho trimestres en los que se presentan registros de casos. Este porcentaje oscila entre el 1 y 8 por ciento, siendo Finlandia el país con menos porcentaje de casos y Republica Checa el de mayor porcentaje. Los periodos en los que se presenta mayor número de contagios son el último trimestre de 2020 y el primero de 2021. Se puede ver también que los países más afectados por la primera oleada han logrado contener la propagación del virus, en comparación con los países que no se vieron mayormente afectados en los primeros trimestres del 2020, tal es el caso de España, Italia y Francia que en ningún trimestre han sobrepasado el 3 por ciento.

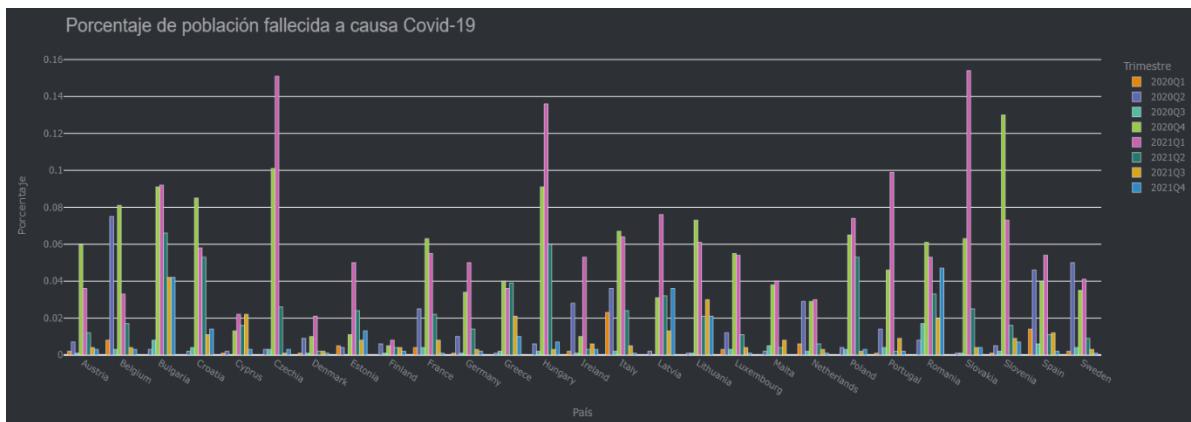


Figura 48. Porcentaje de población Fallecida

El grafico anterior muestra el porcentaje de población fallecida por Covid-19 en cada uno de los 27 países de la UE; este porcentaje oscila entre 0.01 y 015 por ciento, siendo Finlandia el país con menos porcentaje de muertes en general durante los ocho periodos; República Checa y Eslovaquia los dos que presentan mayor registro de muertes durante un trimestre, siendo para ambos casos el primero de 2021. Al igual que con los casos positivos, los picos se encuentran en los trimestres 4 de 2020 y primero de 2021, en ambos casos se percibe un notorio descenso consecuente con el inicio de periodo de vacunación

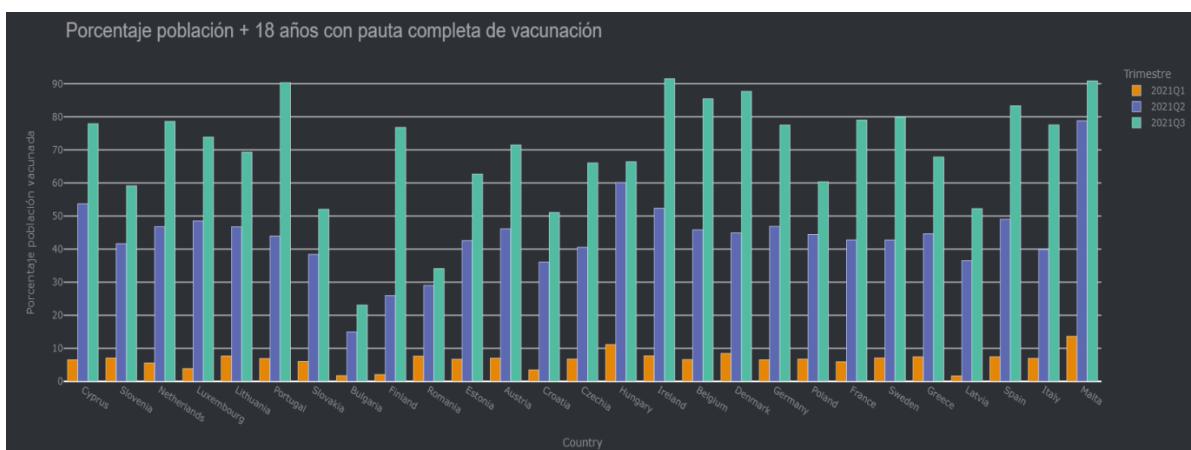


Figura 49. Porcentaje acumulado de población mayor de 18 años con la pauta completa de vacunación

El gráfico anterior muestra el porcentaje acumulado de población mayor de 18 años con la pauta completa de vacunación siendo para la vacuna Janssen una sola dosis y para el resto de vacunas dos dosis, excluyendo pautas de refuerzo. Durante el primer trimestre se puede ver cierta equidad en el proceso de vacunación, la mayoría de países rozando el 10%, los 3 países con menos dosis aplicadas en el primer trimestre son Finlandia, Bulgaria y Letonia y los dos que sobrepasaron el 10% Malta y Hungría. En los períodos posteriores se pierde la armonía inicial, puesto que en el tercer trimestre de vacunación se perciben fuertes diferencias entre países. Finlandia, que inició con un ratio bajo, al finalizar el tercer trimestre se acerca al 80%

de población con pauta completa de vacunación, mientras que Bulgaria permanece con porcentajes Bajos siendo junto a Rumania los países de la UE con menos porcentaje de población vacunada, no alcanzando el 40%. Son tres los países con mas población vacunada, llegando al 90%, Portugal, Malta e Irlanda

PIB por industria

Siendo el objetivo de este trabajo conocer como el Covid-19 ha afectado las economías de los países de la Unión Europea se hace necesario conocer las principales economías que sustentan a cada uno de los países, para esto se identifican las tres economías que más aportación tienen sobre el PIB en cada uno de los países. NACE (Nomenclatura de actividades económicas) es la clasificación estadística europea de actividades económicas; agrupa a las organizaciones de acuerdo con sus actividades comerciales. Para este caso, tomando la división de Eurostat para este Data set, puesto que en otros se profundiza más, con sub divisiones, se tiene la siguiente clasificación general:

Cód. NACE	Descripción
C	Manufacturing
G-I	Wholesale and retail trade, transport, accommodation and food service activities
O-Q	Public administration, defence, education, human health and social work activities
K	Financial and insurance activities
B-E	Industry (except construction)
L	Real estate activities
M_N	Professional, scientific and technical activities; administrative and support service activities
A	Agriculture, forestry and fishing
F	Construction
J	Information and communication
R-U	Public administration, defence, education, human health and social work activities

Tabla 7. Actividades económicas NACE

En los siguientes gráficos sunburst se representa el ranking de Actividades en el primero el Top 1 y los países para los que la actividad en concreto tiene mayor porcentaje de aportación al PIB, el segundo el Top 2 y los países para los que actividad tiene la segunda mayor aportación, así mismo con el tercero.

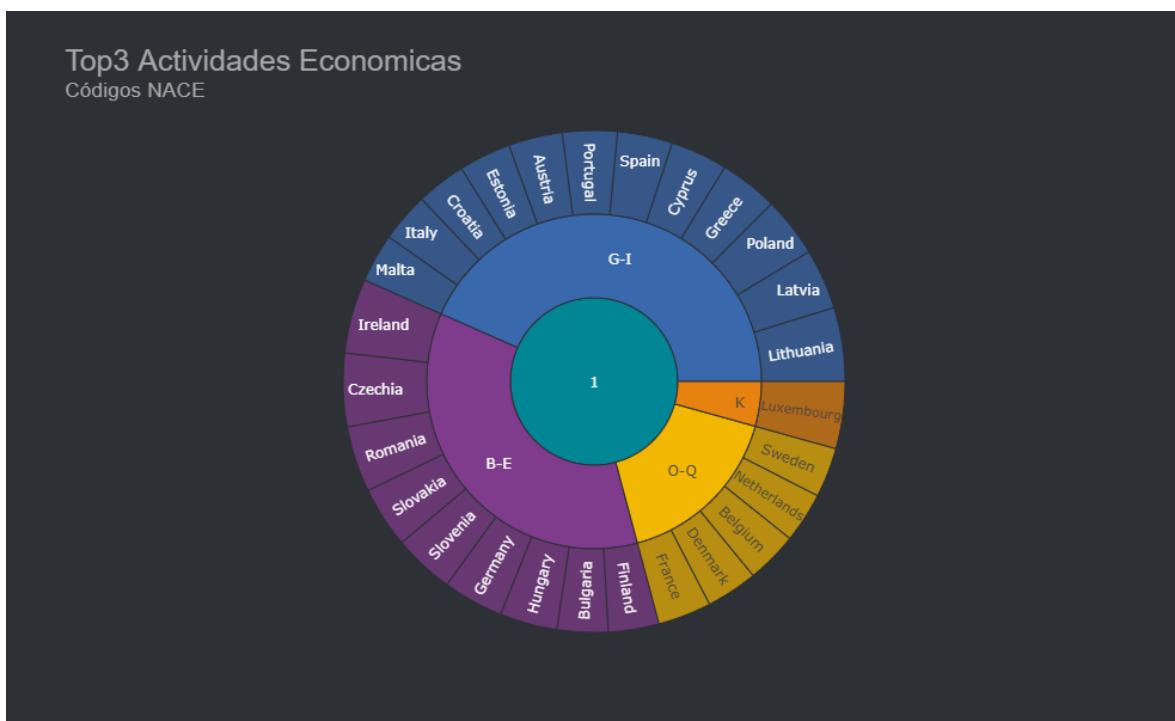


Figura 50. Sunburst Top 1 Actividades económicas

Doce países tienen como principal actividad económica aportadora al PIB, G-I, Comercio al por mayor y al por menor; transporte, almacenamiento y Hostelería y *solo uno K* Actividades financieras y de seguros

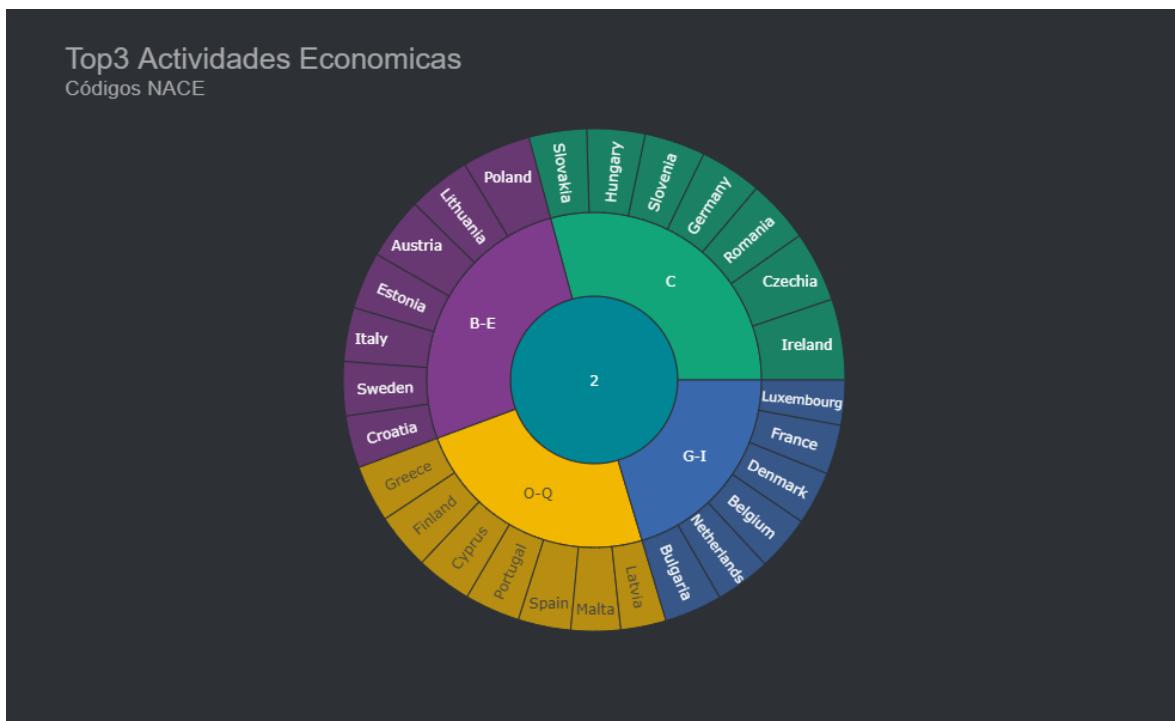


Figura 51. Sunburst Top 2 Actividades económicas

En la segunda posición se repiten las actividades G-I, B-E, O-Q y se incluye C, Industria manufacturera, siendo la segunda actividad económica aportadora al PIB de 7 países.

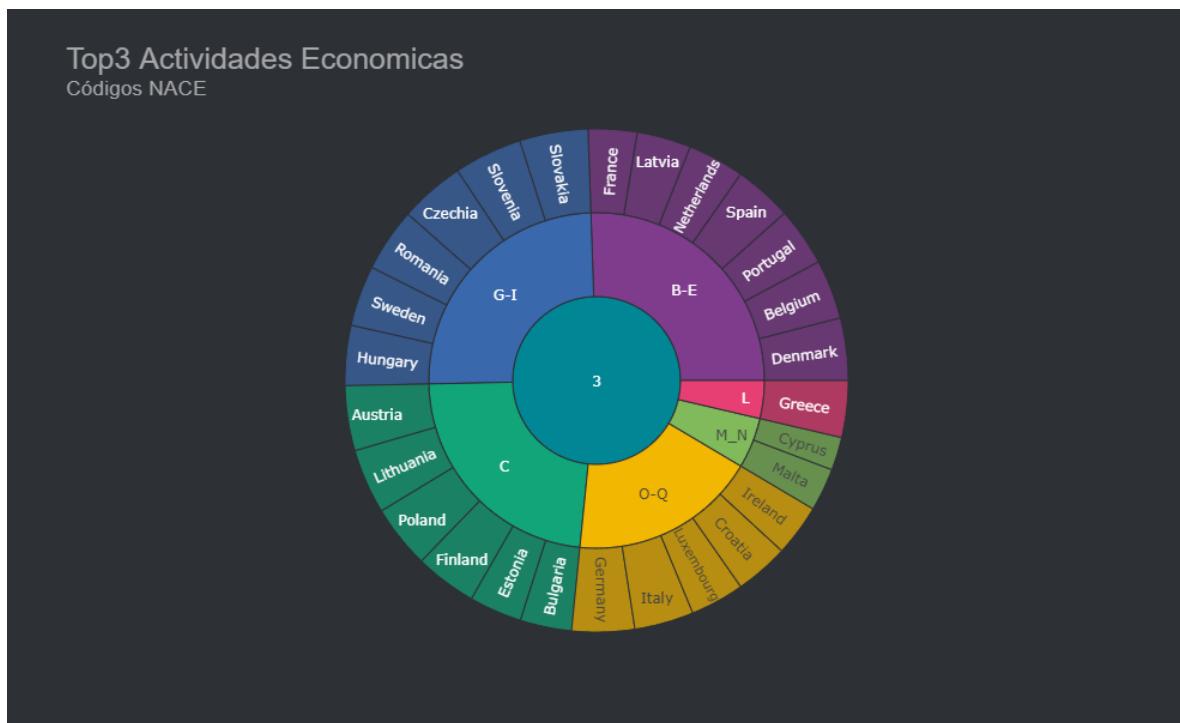


Figura 52. Sunburst Top 3 Actividades económicas

En la tercera posición se repiten las 4 actividades anteriores y se incluyen, M_N, Actividades profesionales, científicas y técnicas; Actividades administrativas y servicios auxiliares como tercera actividad aportadora para Chipre y Malta y L, Actividades inmobiliarias para Grecia

Del total de 12 actividades son 7 las que primen entre los países de la UE:

- C, Industria manufacturera
- G-I, Comercio al por mayor y al por menor; transporte, almacenamiento y Hostelería
- O-Q, Administración pública y defensa; Seguridad Social obligatoria, Educación, Actividades sanitarias y de servicios sociales
- K, Actividades financieras y de seguros
- B-E, Industria (excepto construcción)
- L, Actividades inmobiliarias
- M_N, Actividades profesionales, científicas y técnicas; Actividades administrativas y servicios auxiliares

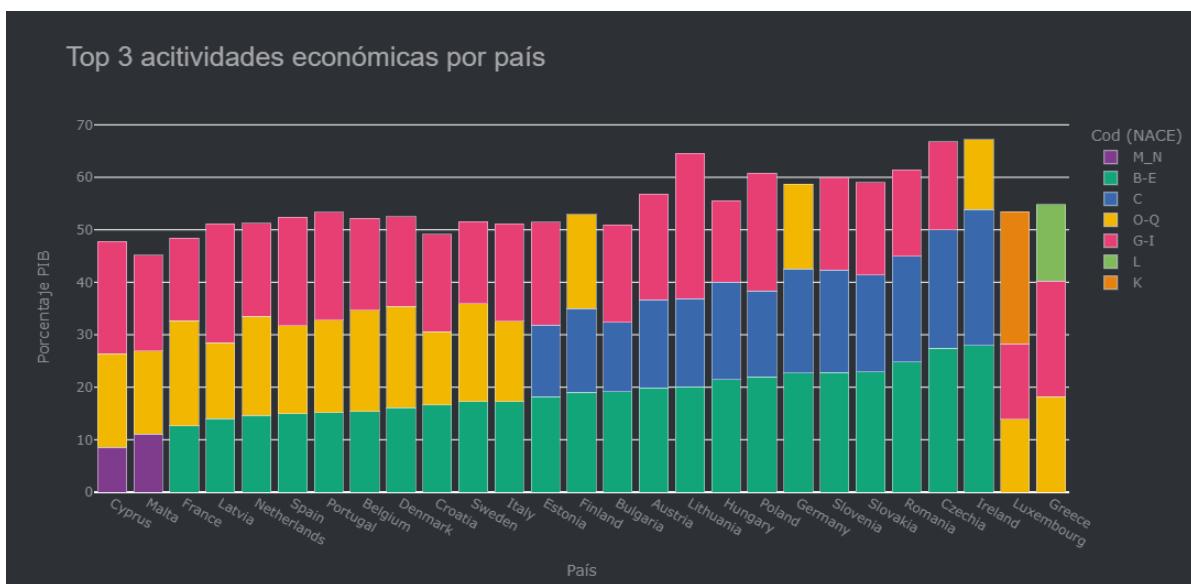


Figura 53. Gráfico de Barras Top 3 Actividades económicas

En el gráfico de barras anterior se puede ver cierta uniformidad en esta distribución de actividades, vemos que, por ejemplo, las actividades económicas que encabezan Francia, Lituania, Holanda, España, Portugal, Bélgica, Dinamarca, Croacia, Suecia e Italia comparten las mismas actividades económicas en cuanto a mayor aportación al PIB; G-I, Comercio al por mayor y al por menor; transporte, almacenamiento y Hostelería está presente en el Top 3 de 24 países; B-E, Industria (excepto construcción) está presente en el Top 3 de 23 países; por otro lado, K, Actividades financieras y de seguros, solo está presente en Luxemburgo y es la que más aporta al PIB; L, Actividades inmobiliarias solo está presente en Grecia y es la tercera con mayor aportación, M_N, Actividades profesionales, científicas y técnicas; Actividades administrativas y servicios auxiliares son la tercera fuerza económica de Malta y Chipre

A partir de la relación de datos anterior puedo dividir los 27 países en 4 bloques, según comparten similitudes en despliegue de actividades económicas:

Bloque 1:

Actividades NACE	Países
O-Q, Administración pública y defensa; Seguridad Social obligatoria, Educación, Actividades sanitarias y de servicios sociales	Irlanda, Finlandia y Alemania
C, Industria manufacturera	
B-E, Industria (excepto construcción)	

Tabla 8. Países bloque 1

Bloque 2:

Actividades NACE	Países
G-I, Comercio al por mayor y al por menor; transporte, almacenamiento y Hostelería	
O-Q, Administración pública y defensa; Seguridad Social obligatoria, Educación, Actividades sanitarias y de servicios sociales	Francia, Lituania, Holanda, España, Portugal, Bélgica, Dinamarca, Croacia, Suecia e Italia
B-E, Industria (excepto construcción)	

Tabla 9. Países bloque 2

Bloque 3:

Actividades NACE	Países
G-I, Comercio al por mayor y al por menor; transporte, almacenamiento y Hostelería	Estonia, Bulgaria, Austria, Lituania, Hungría, Polonia, Eslovenia, Eslovaquia, Romania y Republica Checa
O-Q, C, Industria manufacturera	
B-E, Industria (excepto construcción)	

Tabla 10. Países bloque 3

Bloque 4:

Actividades NACE	Países
G-I, Comercio al por mayor y al por menor; transporte, almacenamiento y Hostelería	
O-Q, Administración pública y defensa; Seguridad Social obligatoria, Educación, Actividades sanitarias y de servicios sociales	Grecia, Luxemburgo, Malta y Chipre
K, Actividades financieras y de seguros	
L, Actividades inmobiliarias	
M_N, Actividades profesionales, científicas y técnicas; Actividades administrativas y servicios auxiliares	

Tabla 11. Países bloque 4

Mercado Laboral

En orden a visualizar y comparar las estadísticas de mercado laboral, tasa de actividad, tasa de ocupación o empleo y tasa de paro o desempleo, usaré los bloques de países expuestos anteriormente seleccionando para cada uno los países más representativos o que han destacado por particularidades en las representaciones anteriores, como es el caso de Finlandia o Bulgaria. En cuanto a la tasa de ocupación por actividad económica expongo en este documento algunos de los países de los bloques mencionados, los gráficos interactivos para cada uno de los 27 países se encuentra en el archivo anexo *Graphics*.

Tasa de Actividad:

El INE Instituto Nacional de estadística de España define la Tasa de actividad como; “el cociente entre el total de activos y la población de 16 y más años” (INE, 2021). Para este caso lo he calculado para ambos sexos y para cada uno de ellos por separado.

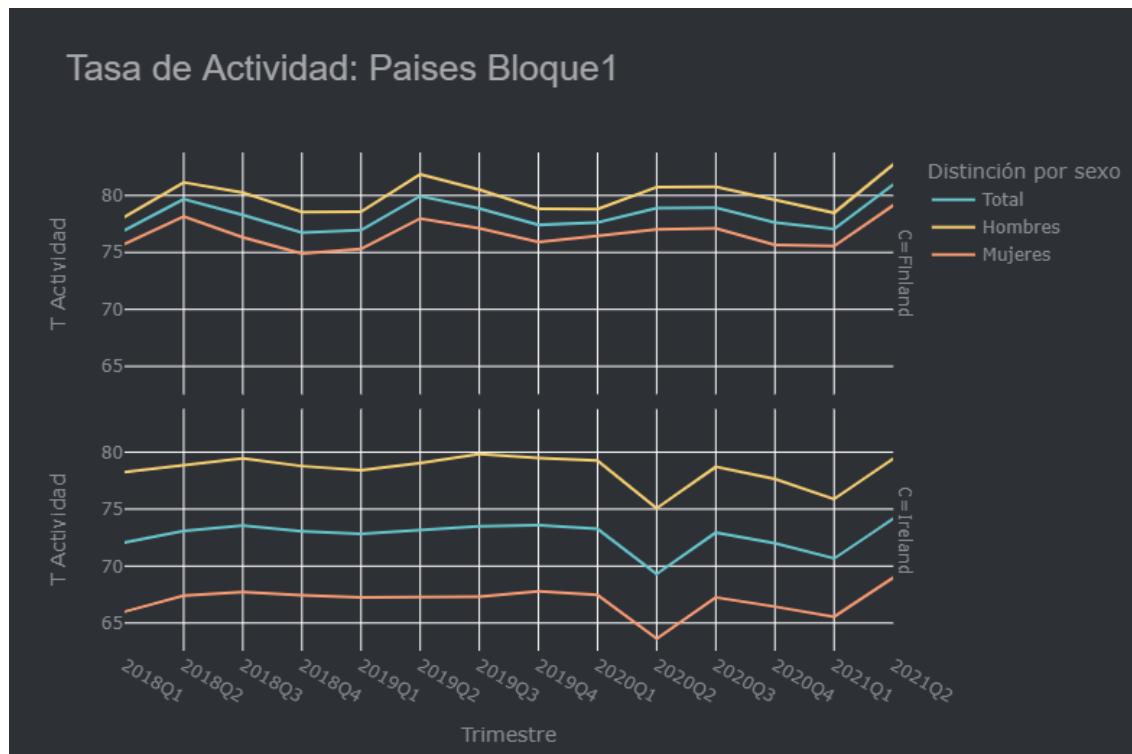


Figura 54. Tasa de Actividad países bloque 1

Para el caso de Irlanda se observa una constante en la Tasa de actividad entre los periodos de 2018 y 2019 llegando al segundo trimestre de 2020 a una caída de 5 puntos para cada distinción, luego se presenta un alza alcanzando los niveles del tercer trimestre de 2019.

Finlandia por su parte llega a los períodos de pandemia con ligeras subidas y bajadas en los períodos de 2018 y 2019 se mantiene relativamente constante en los períodos de 2020 y 2021 con una ligera subida en el segundo trimestre de 2021

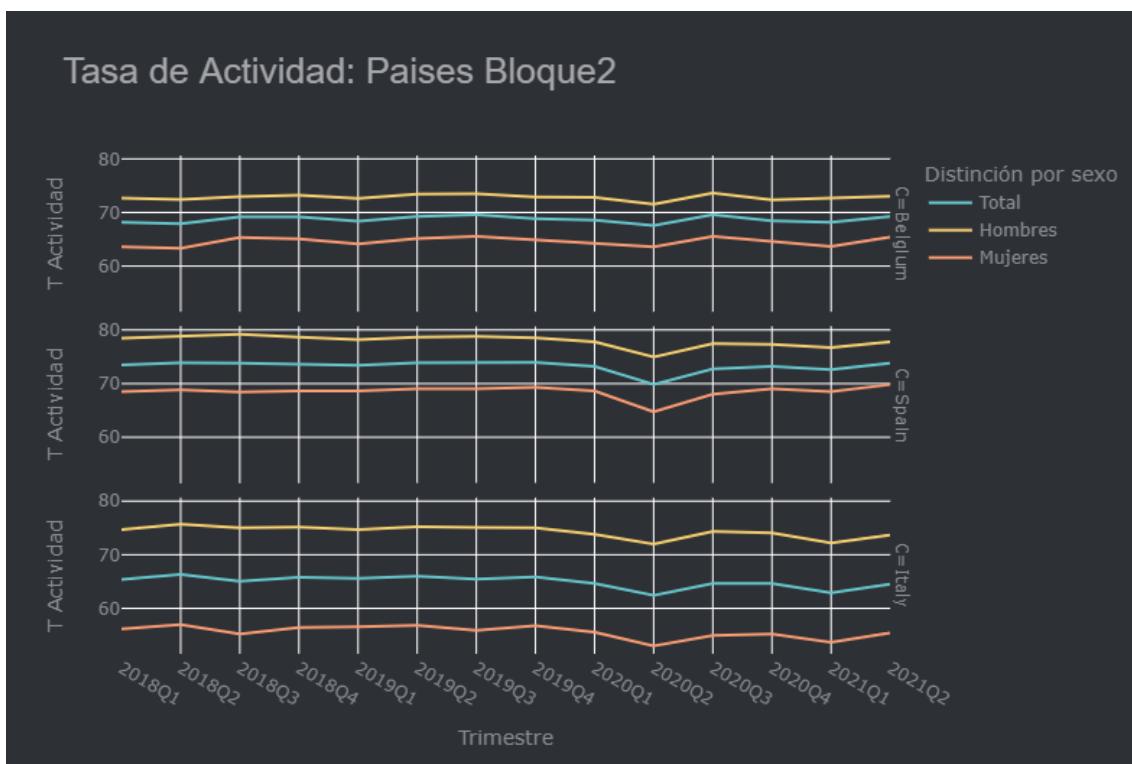


Figura 55. Tasa de Actividad países bloque 2

Para el caso de los países representativos del Bloque dos, los tres presentan una caída en la Tasa de actividad en el segundo trimestre de 2020, retomando los niveles de 2018 y 2019 en el segundo trimestre de 2021

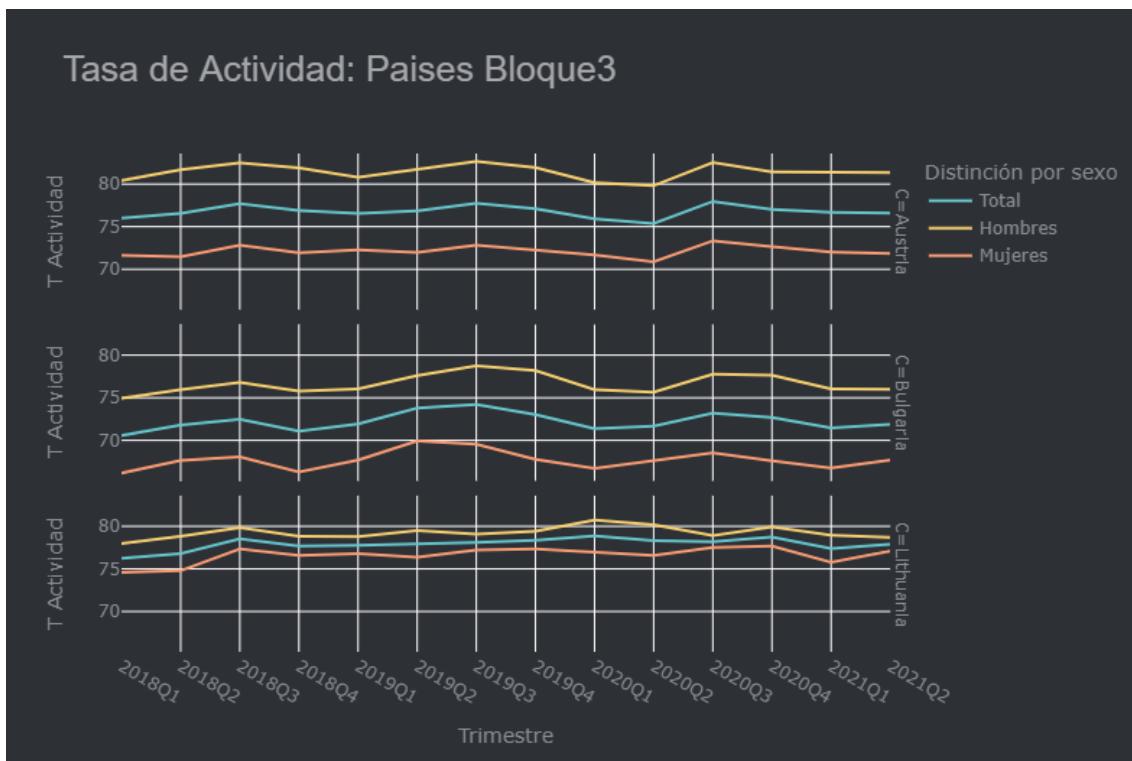


Figura 56. Tasa de Actividad países bloque 3

En el caso de los países representativos del Bloque 3, Austria y Bulgaria presentan caídas en la Tasa de actividad durante los dos primeros trimestres de 2020, mientras que Lituania se mantiene constante.

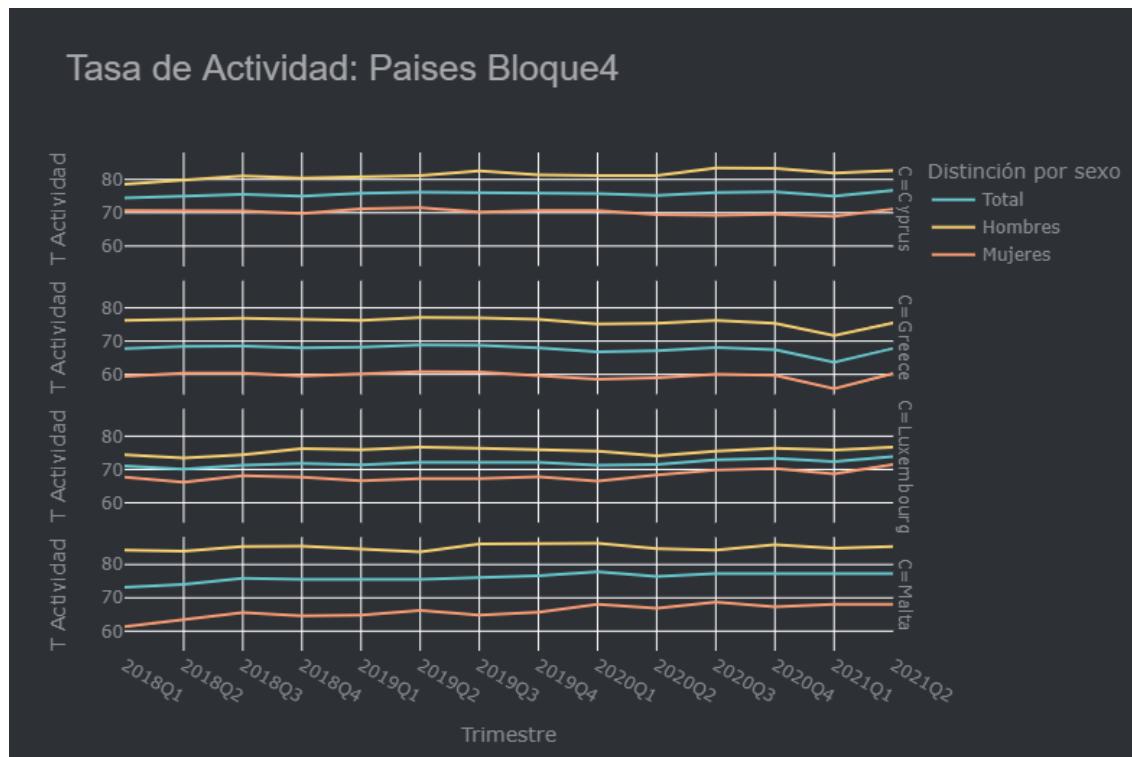


Figura 57. Tasa de Actividad países bloque 4

En el Bloque de países 4, Grecia es el que presenta caídas en la Tasa de actividad, esta vez para el primer trimestre de 2021

De los anteriores Gráficos se deduce:

- + La tasa de actividad se encuentra en general entre 65 y 80%
- + Los países con Tasa de Actividad más alta son Finlandia y Lituania, por encima del 75%
- + Los que presentan caídas considerables en los trimestres de pandemia son Irlanda, España Italia y Grecia
- + La tasa de ocupación es más baja en mujeres, siendo los que tienen una brecha más alta Irlanda, Italia, Austria y Bulgaria; y la brecha más baja Finlandia, Lituania y Luxemburgo

Tasa de Ocupación:

La tasa de ocupación o empleo la define el INE como; ‘el cociente entre el total de ocupados y la población de 16 y más años’ (INE, 2021). Para este caso lo he calculado para ambos sexos y para cada uno de ellos por separado. Con esta se quiere medir como ha variado el empleo entre los periodos inmediatamente anteriores a la pandemia con el periodo de pandemia.

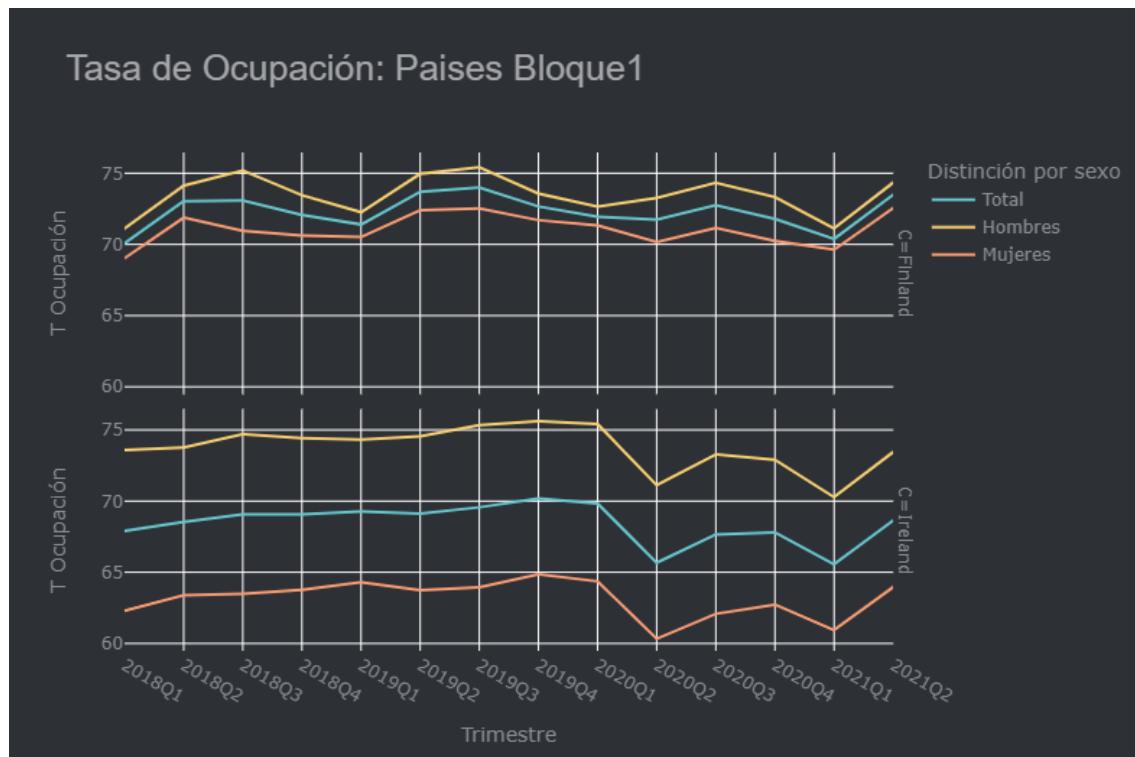


Figura 58. Tasa de Ocupación países bloque 1

Para el caso de Irlanda se observa una constante en la Tasa de ocupación entre los periodos de 2018 y 2019 llegando al segundo trimestre de 2020 a una caída de 5 puntos de la que aún no logra recuperarse pues para el segundo trimestre de 2021 no logra llegar a los niveles de 2018 y 2019.

Finlandia por su parte llega a los periodos de pandemia, al igual que con la tasa de actividad, con ligeras subidas y bajadas en los periodos de 2018 y 2019 `presenta la mayor caída en el primer trimestre de 2021 aunque no baja más de lo que llegó a estar en el primer trimestre de 2018

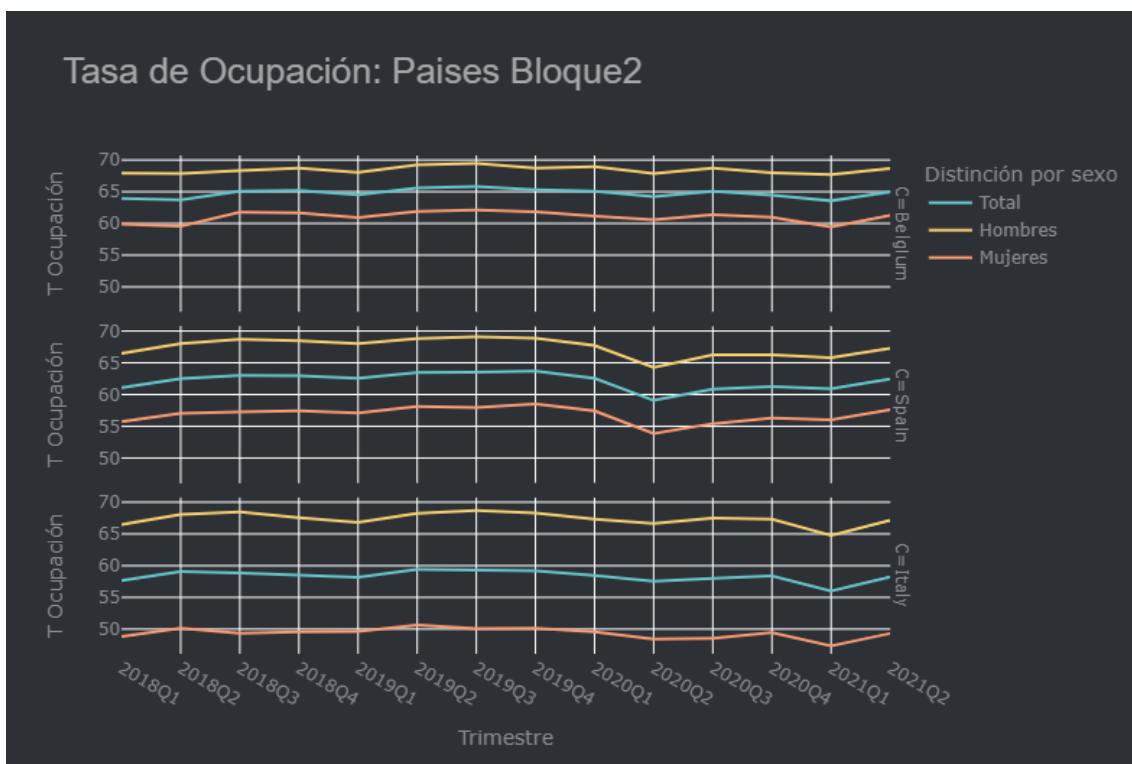


Figura 59. Tasa de Ocupación países bloque 2

Para el caso de los países seleccionados del Bloque 2, Bélgica se mantiene constante, España presenta una caída de aproximadamente 5 puntos en el segundo trimestre de 2020, mientras que Italia presenta una leve bajada en el primer trimestre de 2021

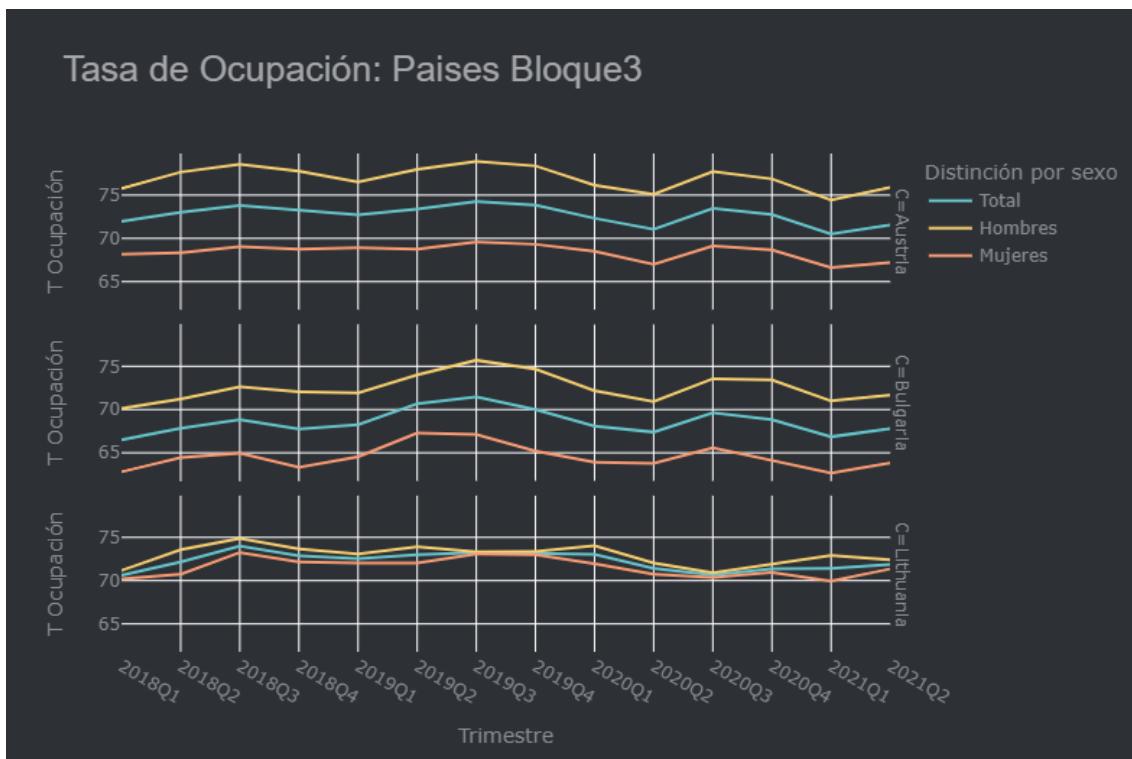


Figura 60. Tasa de Ocupación países bloque 3

Al igual que con la Tasa de actividad, en el caso de los países representativos del Bloque 3, Austria y Bulgaria presentan caídas en la Tasa de ocupación durante los dos primeros trimestres de 2020, mientras que Lituania se sigue manteniendo constante.

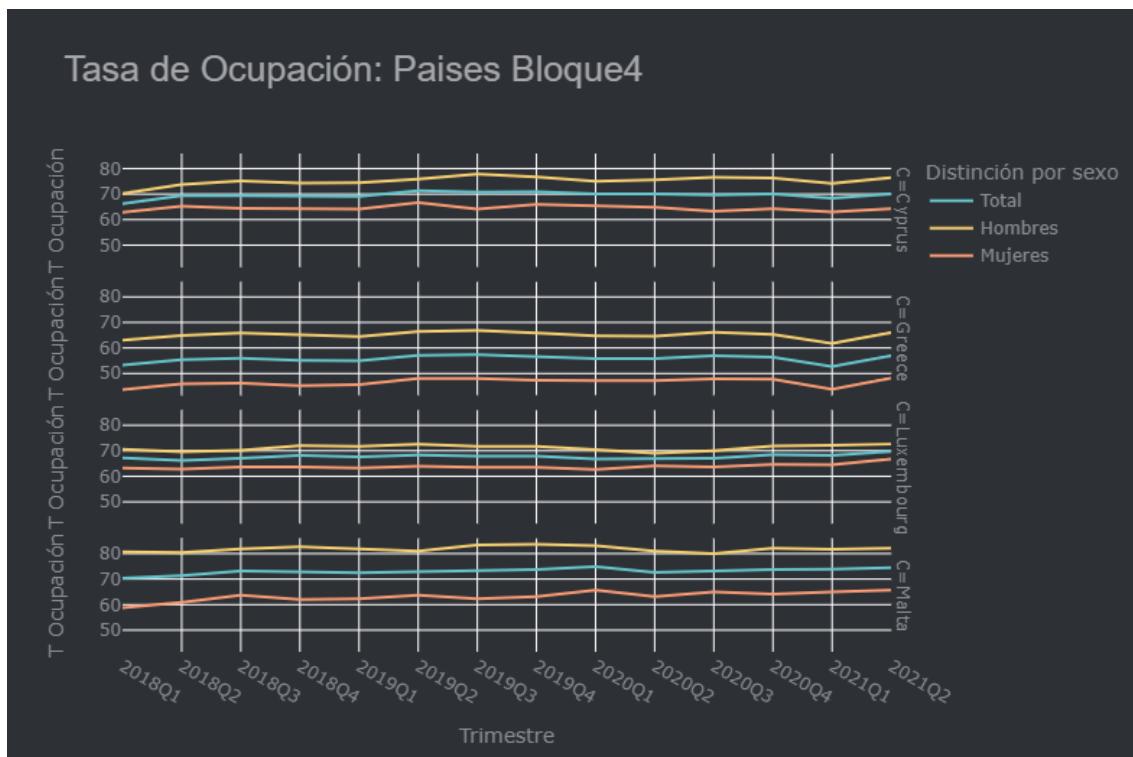


Figura 61. Tasa de Ocupación países bloque 4

En el Bloque de países 4, sigue siendo Grecia quien presenta caídas esta vez en la Tasa de ocupación , para el primer trimestre de 2021

De los anteriores gráficos se deduce:

- + La tasa de actividad se encuentra en general entre 50 y 75%
- + Los países de los bloques 1, 2, y 3 tienen caídas en la tasa de ocupación en mayor o menor medida en los trimestres de pandemia, sin embargo los países del bloque 4 a excepción de Grecia, es decir Chipre, Luxemburgo y Malta no presentan mayores alteraciones, se mantienen constantes
- + La brecha entre hombres y mujeres se sigue presentando, siendo en general las mujeres las que tienen tasa de ocupación más baja, la brecha más alta la tienen Irlanda, Italia y Austria

Tasa de Paro:

La tasa de ocupación o empleo la define el INE como ‘el cociente entre el número de parados y el de activos. Se calcula para ambos sexos y para cada uno de ellos por separado’ (INE, 2021). Con esta se quiere medir como ha variado el desempleo entre los periodos inmediatamente anteriores a la pandemia con el periodo de pandemia

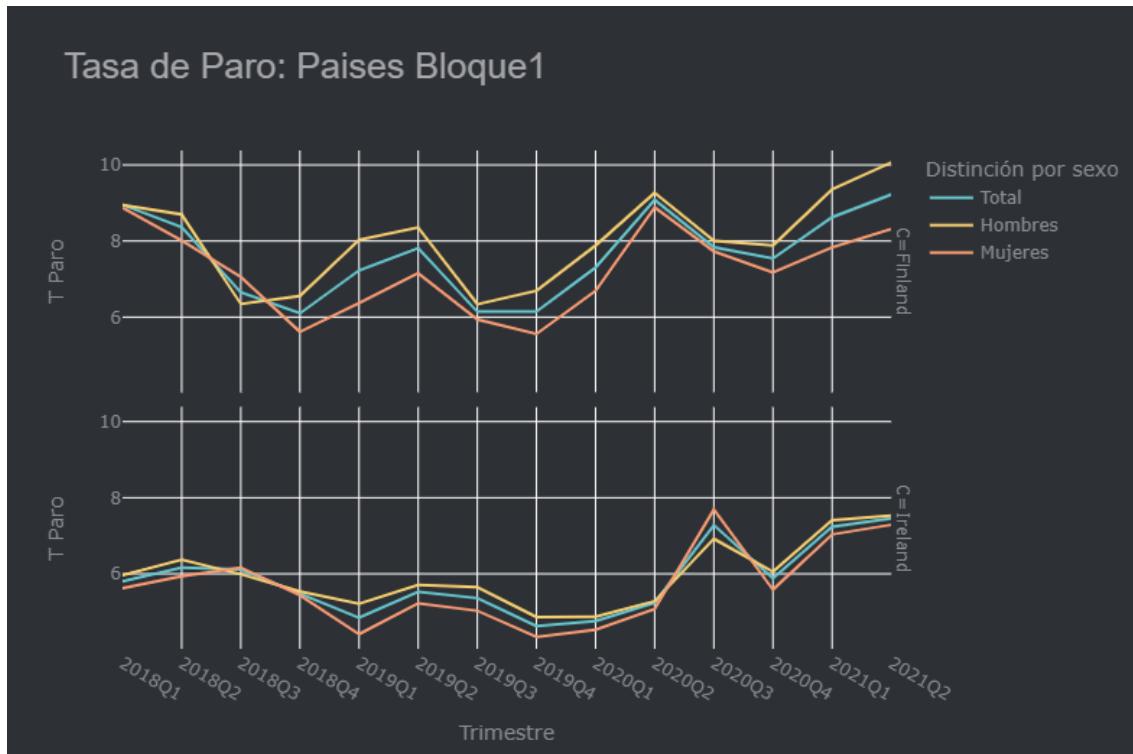


Figura 62. Tasa de Paro países bloque 1

A diferencia de la Tasa de actividad y de ocupación, para los casos de Finlandia e Irlanda, la Tasa de paro alcanza su peor momento en el segundo trimestre de 2021, llegando a niveles que no se presentaron en los periodos de 2018 y 2019

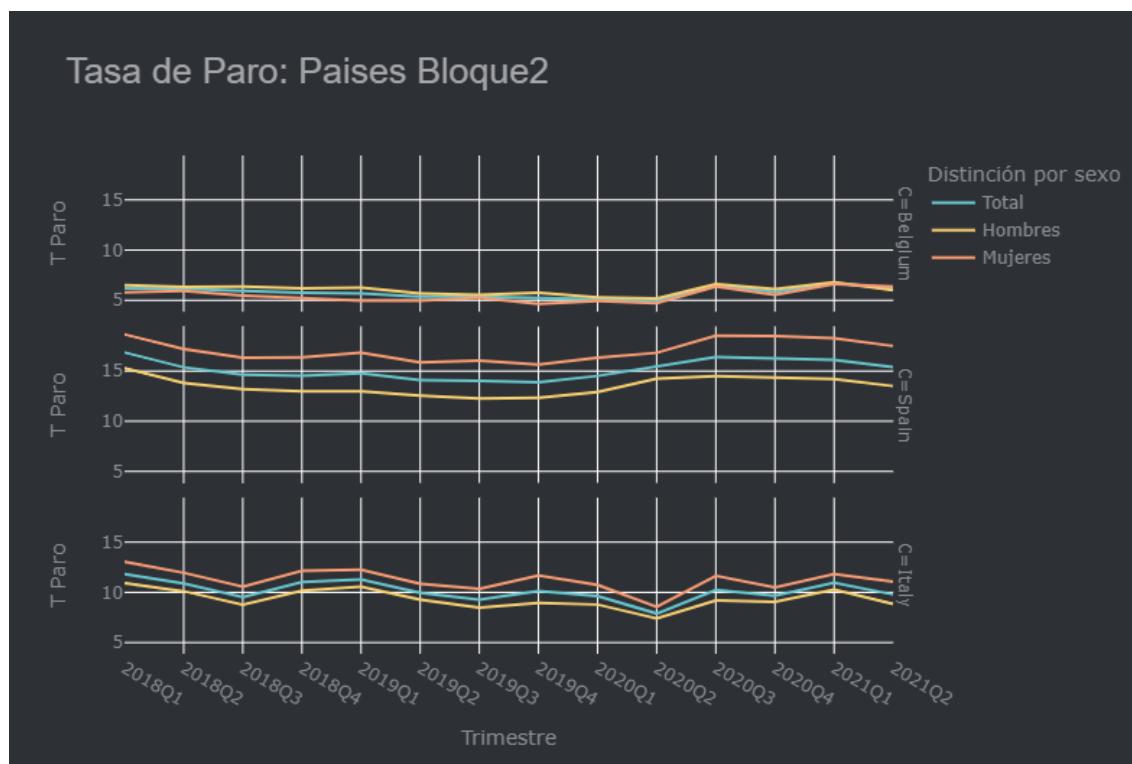


Figura 63. Tasa de Paro países bloque 2

Para el caso de los países seleccionados del Bloque 2, se presentan subidas en la Tasa de paro durante los trimestres 2, 3 y 4 de 2020, siendo en España ligeramente más alta

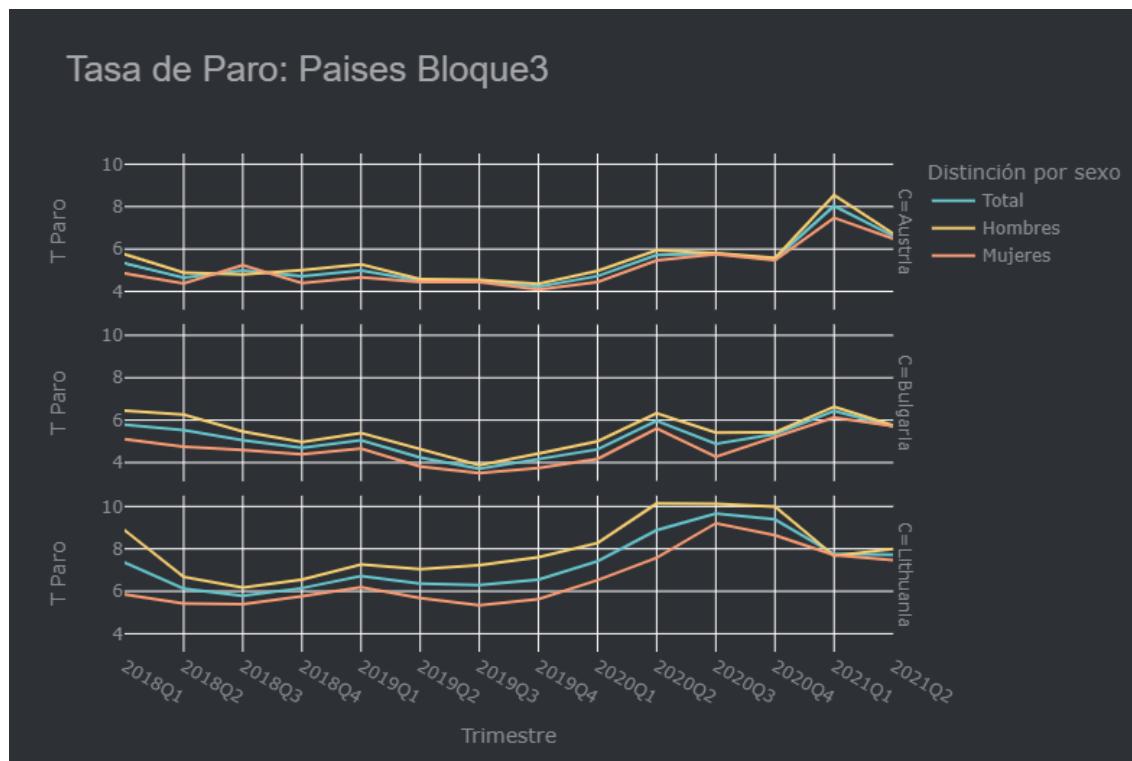


Figura 64. Tasa de Paro países bloque 3

En el caso de los países representativos del Bloque 3, Austria es el que presenta una fuerte subida en la Tasa de paro, llegando a su mayor punto en el primer trimestre de 2021, 4 puntos por encima de la tasa que llevaba en el último trimestre de 2019, sube del 4 a 8 por ciento, Lituania también presenta una alza importante pasando de 6 por ciento en el tercer trimestre de 2019 a 10 por ciento en el primer trimestre de 2020 manteniéndose así durante todo el año 2020, igualmente Bulgaria presenta subidas llegando al mayor punto en el primer trimestre de 2021

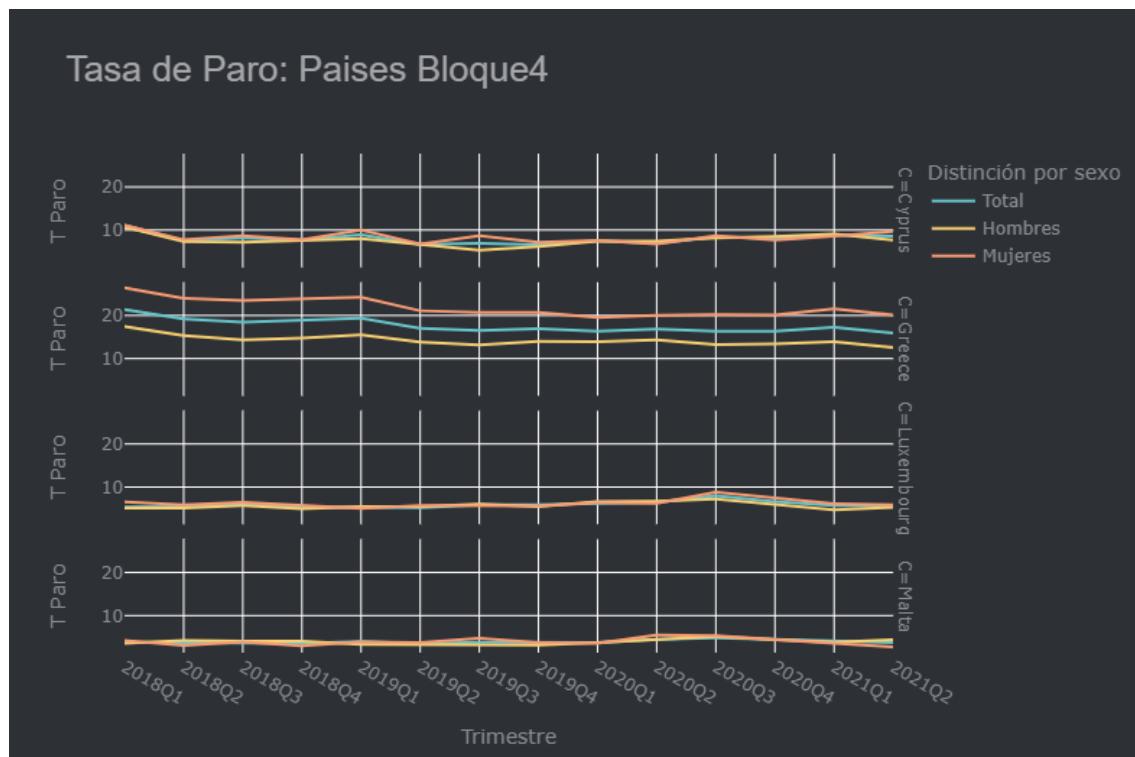


Figura 65. Tasa de Paro países bloque 4

De los gráficos anteriores se deduce:

- + La tasa de paro en general se encuentra entre el 5 y 10%, con excepción de dos países, Grecia que ronda el 20% y España que ronda el 15%
- + Los países que presentan mayor aumento durante los meses de pandemia son Finlandia, Irlanda, Austria, Bulgaria y Lituana
- + La tasa de paro entre hombres y mujeres en general se mantiene en los mismo porcentajes, en España y Grecia es más alta para las mujeres y en Austria, Lituania y Finlandia es mayor en los hombres

Tasa de Ocupación por Actividad económica:

La tasa de ocupación o empleo por Actividad económica, se ha calculado entre el total de ocupados para cada actividad económica, es decir el número de personas empleadas bajo cada categoría NACE y la población entre 16 y 64 años

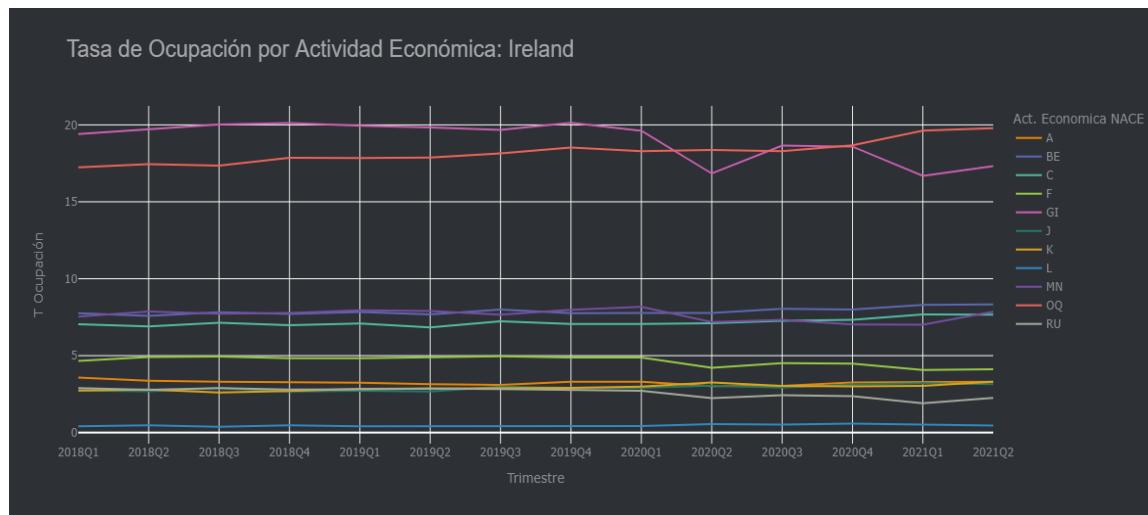


Figura 66. Tasa de Ocupación por Actividad económica Irlanda

En el caso de Irlanda, las dos actividades NACE que más generan empleo son G-I Comercio al por mayor y al por menor; transporte, almacenamiento y Hostelería, representada en color rosa y O-Q Administración pública y defensa; Seguridad Social obligatoria, Educación, Actividades sanitarias y de servicios sociales, representada con color rojo. En los periodos anteriores a la pandemia G-I ocupa el primer puesto, con la llegada de la pandemia O-Q pasa a ser la actividad económica con mayor tasa de empleo.

Curiosamente G-I en Irlanda es una de las dos actividades económicas con mayor Tasa de ocupación, pero no se encuentra entre las tres con mayor aportación al PIB

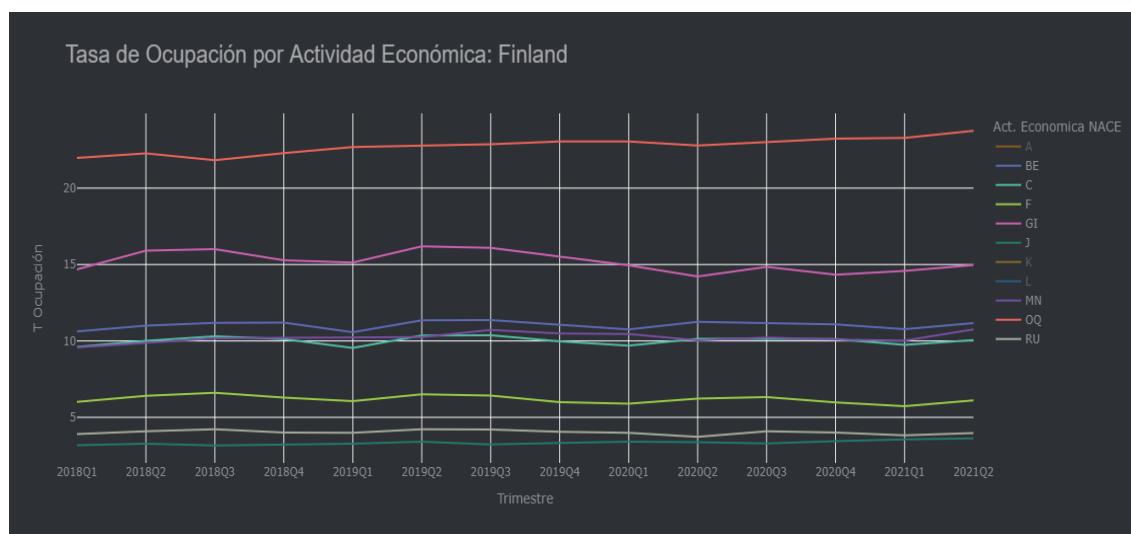


Figura 67. Tasa de Ocupación por Actividad económica Finlandia

En el caso de Finlandia, las dos actividades NACE que más generan empleo son O-Q Administración pública y defensa; Seguridad Social obligatoria, Educación, Actividades sanitarias y de servicios sociales, representada con color rojo; y G-I Comercio al por mayor y al por menor; transporte, almacenamiento y Hostelería, representada en color rosa.

Para O-Q no se perciben mayores variaciones en cuanto a la tasa de ocupación durante los periodos de pandemia, no es el caso de G-I, que presenta bajadas durante el segundo y cuarto trimestre de 2020; no obstante O-Q se mantiene en primer lugar y G-I en segundo.

En este caso, para Finlandia G-I también es una de las dos actividades económicas con mayor Tasa de ocupación, pero no se encuentra entre las tres con mayor aportación al PIB

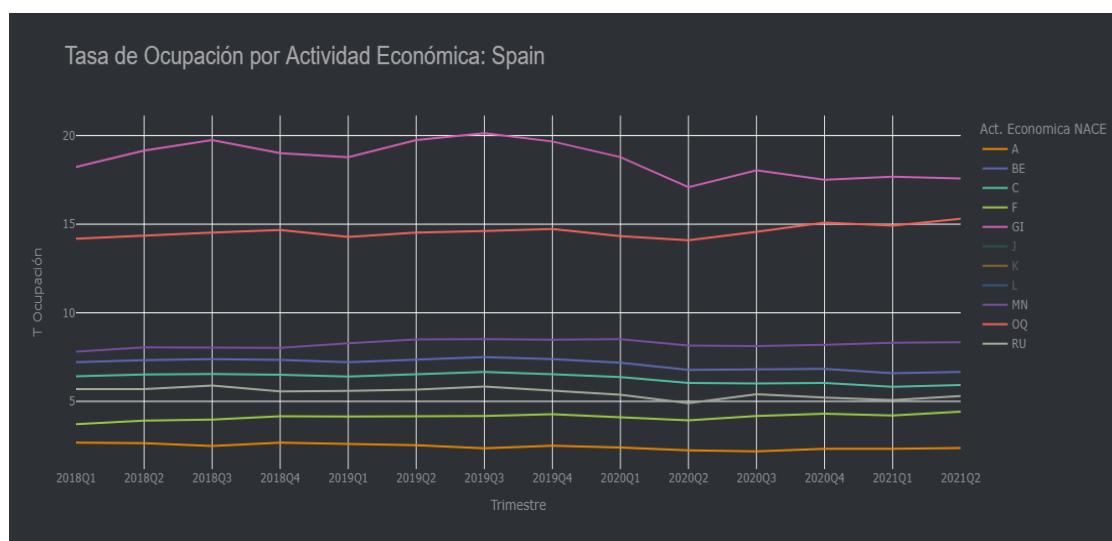


Figura 68. Tasa de Ocupación por Actividad económica España

Para España, las dos actividades NACE que más generan empleo son igualmente G-I Comercio al por mayor y al por menor; transporte, almacenamiento y Hostelería, representada en color rosa y O-Q Administración pública y defensa; Seguridad Social obligatoria, Educación, Actividades sanitarias y de servicios sociales, representada con color rojo. G-I presenta una fuerte caída en la tasa de Ocupación durante el segundo trimestre de 2020, una ligera subida en el tercer trimestre de 2020, estabilizándose en los siguientes trimestres, pero manteniéndose por debajo en comparación con los períodos 2018 y 2020. O-Q se mantiene constante hasta el segundo trimestre de 2020 y para subir a partir del tercer trimestre de 2020, acercándose al porcentaje de G-I.

En este caso las dos actividades que más generan empleo coinciden con dos de las tres que más aportación hacen al PIB.

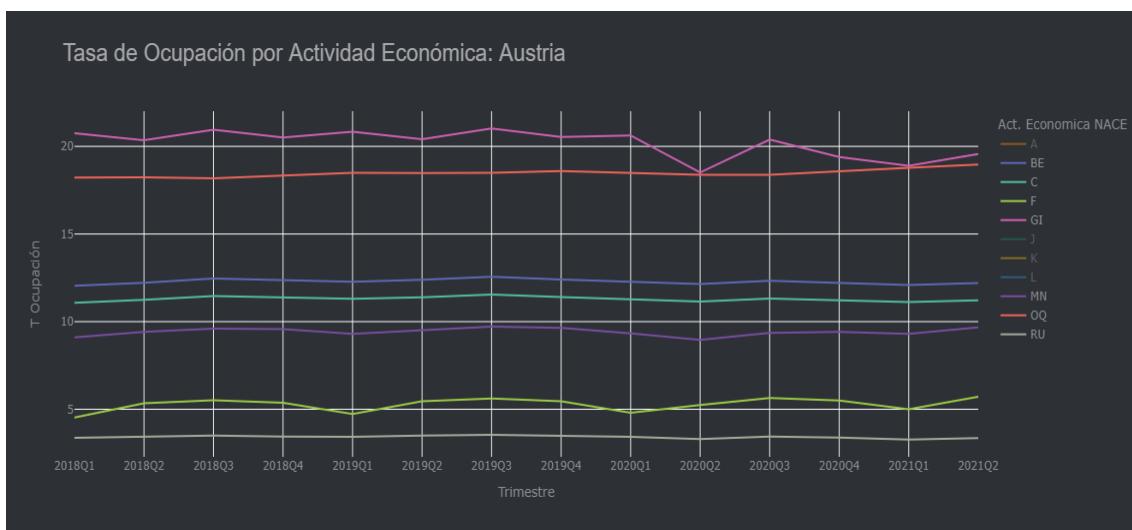


Figura 69. Tasa de Ocupación por Actividad económica Austria

Igualmente, en Austria, las dos actividades NACE que más generan empleo son G-I Comercio al por mayor y al por menor; transporte, almacenamiento y Hostelería, representada en color rosa y O-Q Administración pública y defensa; Seguridad Social obligatoria, Educación, Actividades sanitarias y de servicios sociales, representada con color rojo. O-Q permanece constante con una muy ligera subida a partir del último trimestre de 2020 y G-I presenta fuertes bajadas en el segundo semestre de 2020 y el primer trimestre de 2021, llegando a estar al mismo nivel de O-Q en estos trimestres puntuales.

En este caso es O-Q una de las actividades que más generan empleo pero que no se encuentran dentro de las tres de mayor aportación al PIB

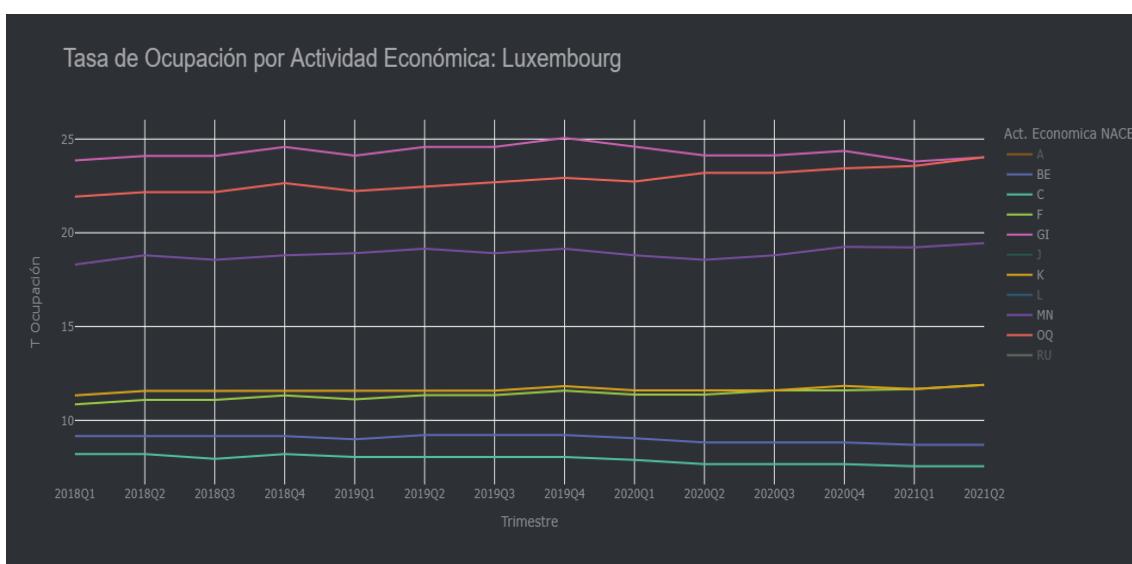


Figura 72. Tasa de Ocupación por Actividad económica Luxemburgo

Finalmente, Luxemburgo sigue un comportamiento similar siendo para este también las dos actividades NACE que más generan empleo son G-I Comercio al por mayor y al por menor; transporte, almacenamiento y Hostelería, representada en color rosa y O-Q Administración pública y defensa; Seguridad Social obligatoria, Educación, Actividades sanitarias y de servicios sociales, representada con color rojo.

En este caso la primera inicia una bajada paulatina, al mismo tiempo que la segunda inicia una subida paulatina a partir del primer trimestre de 2020 llegándose a equiparar en el segundo trimestre 2021.

6. Conclusiones

- IV. La actividad económica NACE, G-I, Comercio al por mayor y al por menor; transporte, almacenamiento y Hostelería, es la que mayor aporta al PIB en la UE en 24 de los 27 países, además es una de las dos que más genera empleo en los 27 países, incluso en aquellos en los que no se encuentra como una de las de mayor aportación al PIB; así mismo es, sin duda, la que más se ha visto afectada a causa del Covid-19, esto se evidencia en las fuertes bajadas de Tasa de Ocupación o empleo que se presentan en general en todos los países de la UE
- V. La segunda actividad NACE O-Q Administración pública y defensa; Seguridad Social obligatoria, Educación, Actividades sanitarias y de servicios sociales, es una de las tres con mayor aportación al PIB en la UE, en 17 de los 27 países, es una de las dos que más genera empleo en los 27 países y al contrario que la G-I, O-Q no se ha visto mayormente afectada por el Covid-19, por el contrario en la mayoría de los casos presenta subidas en la Tasa de Ocupación o empleo.
- VI. Se puede observar una relación proporcional en cuanto a porcentaje de casos positivos y muertes por Covid-19 y efectos en el mercado laboral; a mayor porcentaje de casos, se presentan más subidas en las Tasas de Paro y bajadas en la tasa de desempleo, tal es el caso de España, Italia y Francia en los que durante el primer y segundo trimestre de 2020 se vieron más afectados por el Covid-19 así mismo se vieron afectadas las estadísticas de mercado laboral durante esos trimestres; mientras que en los países en los que se han presentado aumento de contagios en los primeros trimestres de 2021 se están viendo afectadas las estadísticas de mercado laboral durante los mismos trimestres, como es el caso de Eslovaquia y Eslovenia, países en los que los casos positivos han aumentado en el último trimestre de 2020 y primer trimestre de 2021 presentan bajadas fuertes en las tasas de ocupación durante el primer trimestre de 2021.
- VII. Así mismo se puede observar una relación proporcional entre el porcentaje de población vacunada y las subidas en las tasas de ocupación, tal es el caso de Portugal, Dinamarca y Luxemburgo quienes tienen porcentajes de vacunación altos y presentan subidas en la tasa de ocupación de los últimos dos trimestres.

7. Referencias bibliográficas

- apache.org. (2020). *Apache Parquet*. Obtenido de <http://parquet.incubator.apache.org/documentation/latest/>
- Cazzaniga, N. (2021). *Eurostat Python Package*. Obtenido de <https://pypi.org/project/eurostat/>
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0*. Obtenido de Step-by-step data mining guid: <https://the-modeling-agency.com/crisp-dm.pdf>
- ECDC. (2021). *Data on COVID-19 vaccination in the EU/EEA*. Obtenido de <https://www.ecdc.europa.eu/en/publications-data/data-covid-19-vaccination-eu-eea>
- ECDC. (2021). *How ECDC collects and processes COVID-19 data*. Obtenido de <https://www.ecdc.europa.eu/en/covid-19/data-collection>
- Ecorys. (2021). *europarl.europa.eu*. Obtenido de [https://www.europarl.europa.eu/RegData/etudes/STUD/2021/662903/IPOL_STU\(2021\)662903_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2021/662903/IPOL_STU(2021)662903_EN.pdf)
- European Comission. (2020). *JRC analyses COVID-19 impact on economy and labour markets to help guide EU response*. Obtenido de <https://ec.europa.eu/jrc/en/news/jrc-analyses-covid-19-impact-economy-and-labour-markets-help-guide-eu-response>
- Eurostat. (2018). *Eurostat La llave de acceso a las estadísticas europeas*. Obtenido de <https://ec.europa.eu/eurostat/documents/4031688/8932118/KS-02-17-839-ES-N.pdf/04b4200a-eba6-477f-8690-b47a308e7e20#:~:text=La%20misi%7B/%7Bo%7D%7Dn%20de%20Eurostat%20es%20proporcionar%20estad%7B/%7Bi%7D%7Dsticas%20de%20alta%20calidad%20para%20Europ>
- IBM. (2021). *CRISP-DM Help Overview*. Obtenido de <https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=dm-crisp-help-overview>
- INE. (2021). *Glosario de Conceptos*. Obtenido de <https://www.ine.es/DEFIne/es/concepto.htm?c=4459&op=30320&p=1&n=20>
- Microsoft. (2021). *What is Power BI?* Obtenido de Power BI get started documentation: <https://docs.microsoft.com/en-us/power-bi/fundamentals/power-bi-overview>
- OMS. (2020). *COVID-19: cronología de la actuación de la OMS*. Obtenido de <https://www.who.int/es/news/item/27-04-2020-who-timeline---covid-19>
- pandas.pydata.org. (2021). Obtenido de <https://pandas.pydata.org/>
- Python.org. (2021). *What is Python? Executive Summary*. Obtenido de <https://www.python.org/doc/essays/blurb/>
- SAS®. (2017). *Introduction to SEMMA*. Obtenido de SAS® Enterprise Miner™ 14.3: Reference Help: <https://documentation.sas.com/doc/en/emref/14.3/n061bzurmej4j3n1jn8bbjjm1a2.htm>
- Sngular. (2021). *CRISP-DM: La metodología para poner orden en los proyectos*. Obtenido de <https://www.sngular.com/es/data-science-crisp-dm-metodologia/>
- Stephan Kudyba. (2014). *Big Data, Mining, and Analytics*. . Obtenido de Transforming Unstructured Data into Useful Information: <https://www.taylorfrancis.com/chapters/edit/10.1201/b16666-14/transforming-unstructured-data-useful-information-meta-brown>

- TIOBE Software BV. (2021). *TIOBE Index for October 2021*. Obtenido de <https://www.tiobe.com/tiobe-index/>
- WHO. (2020). *Información básica sobre la COVID-19*. Obtenido de <https://www.who.int/es/news-room/q-a-detail/coronavirus-disease-covid-19>

8. Anexos

8.1. Diccionario variable unit Eurostat

Unit Code	Unit Description
CLV15_MEUR	Chain linked volumes (2015), million euro
CLV05_MNAC	Chain linked volumes (2005), million units of national currency
PD_PCH_PRE_EUR	Price index (implicit deflator), percentage change on previous period, euro
CLV15_MNAC	Chain linked volumes (2015), million units of national currency
PD05_NAC	Price index (implicit deflator), 2005=100, national currency
PD10_EUR	Price index (implicit deflator), 2010=100, euro
PD15_EUR	Price index (implicit deflator), 2015=100, euro
CLV_I10	Chain linked volumes, index 2010=100
PD_PCH_PRE_NAC	Price index (implicit deflator), percentage change on previous period, national currency
PYP_MNAC	Previous year prices, million units of national currency
PD15_NAC	Price index (implicit deflator), 2015=100, national currency
CLV10_MNAC	Chain linked volumes (2010), million units of national currency
PC_GDP	Percentage of gross domestic product (GDP)
CP_MEUR	Current prices, million euro
PD10_NAC	Price index (implicit deflator), 2010=100, national currency
CON_PPCH_PRE	Contribution to GDP growth, percentage point change on previous period
PD_PCH_SM_EUR	Price index (implicit deflator), percentage change compared to same period in previous year, euro
CLV05_MEUR	Chain linked volumes (2005), million euro
PD_PCH_SM_NAC	Price index (implicit deflator), percentage change compared to same period in previous year, national currency
CLV_PCH_PRE	Chain linked volumes, percentage change on previous period
CLV_I05	Chain linked volumes, index 2005=100
CLV10_MEUR	Chain linked volumes (2010), million euro
PYP_MEUR	Previous year prices, million euro
CLV_PCH_SM	Chain linked volumes, percentage change compared to same period in previous year
PD05_EUR	Price index (implicit deflator), 2005=100, euro
CLV_I15	Chain linked volumes, index 2015=100
CP_MNAC	Current prices, million units of national currency
CON_PPCH_SM	Contribution to GDP growth, percentage point change compared to same period in previous year
CLV_PCH_ANN	Chain linked volumes, annualized percentage change on previous
PCH_SM_PER	Percentage change compared to same period in previous year (based on persons)
THS_PER	Thousand persons
PCH_PRE_PER	Percentage change on previous period (based on persons)
THS	Thousand

8.2. Diccionario variable s_adj Eurostat

s_adj Code	s_adj Description
NSA	Unadjusted data (i.e. neither seasonally adjusted nor calendar adjusted data)
SA	Seasonally adjusted data, not calendar adjusted data
CA	Calendar adjusted data, not seasonally adjusted data
SCA	Seasonally and calendar adjusted data

8.3. Diccionario variable na_item Eurostat

na_item Code	na_item Description
B111	External balance - Goods
P3_P5	Final consumption expenditure and gross capital formation
P62	Exports of services
P72	Imports of services
P52_P53	Changes in inventories and acquisitions less disposals of valuables
D21	Taxes on products
P3_S13	Final consumption expenditure of general government
YA0	Statistical discrepancy (expenditure approach)
P31_S13	Individual consumption expenditure of general government
P61	Exports of goods
YA1	Statistical discrepancy (production approach)
P32_S13	Collective consumption expenditure of general government
P71	Imports of goods
P7	Imports of goods and services
D1	Compensation of employees
D2X3	Taxes on production and imports less subsidies
P53	Acquisitions less disposals of valuables
D3	Subsidies
B1G	Value added, gross
D12	Employers' social contributions
P3	Final consumption expenditure
D31	Subsidies on products
P31_S14_S15	Household and NPISH final consumption expenditure
P51G	Gross fixed capital formation
D21X31	Taxes less subsidies on products
P31_S15	Final consumption expenditure of NPISH
YA2	Statistical discrepancy (income approach)

P5G	Gross capital formation
P41	Actual individual consumption
D2	Taxes on production and imports
B1GQ	Gross domestic product at market prices
B112	External balance - Services
P52	Changes in inventories
P3_P6	Final consumption expenditure, gross capital and exports of goods and services
P31_S14	Final consumption expenditure of households
P6	Exports of goods and services
D11	Wages and salaries
B2A3G	Operating surplus and mixed income, gross
B11	External balance of goods and services
D1	Compensation of employees
B1G	Value added, gross
D12	Employers' social contributions
D11	Wages and salaries
POP_NC	Total population national concept
SAL_NC	Employees national concept
SELF_DC	Self-employed domestic concept
SAL_DC	Employees domestic concept
EMP_NC	Total employment national concept
SELF_NC	Self-employed national concept
EMP_DC	Total employment domestic concept

8.4. Diccionario variable nace_r2 Eurostat

nace_r2 Code	nace_r2 Description
K	Financial and insurance activities
F	Construction
L	Real estate activities
C	Manufacturing
J	Information and communication
TOTAL	Total - all NACE activities
M_N	Professional, scientific and technical activities; administrative and support service activities
A	Agriculture, forestry and fishing
R-U	Arts, entertainment and recreation; other service activities; activities of household and extra-territorial organizations and bodies
O-Q	Public administration, defence, education, human health and social work activities
G-I	Wholesale and retail trade, transport, accommodation and food service activities
B-E	Industry (except construction)

8.5. Diccionario variable citizen Eurostat

citizen Code	citizen Description
NEU28_FOR	Non-EU28 countries (2013-2020) nor reporting country
NRP	No response
EU28_FOR	EU28 countries (2013-2020) except reporting country
STLS	Stateless
EU27_2020_FOR	EU27 countries (from 2020) except reporting country
NEU27_2020_FOR	Non-EU27 countries (from 2020) nor reporting country
TOTAL	Total
EU15_FOR	EU15 countries (1995-2004) except reporting country
NEU15_FOR	Non-EU15 countries (1995-2004) nor reporting country
FOR	Foreign country
NAT	Reporting country

8.6 Graphics.html

Archivo adjunto a este documento.