Entregable 2: Preproceso, Transformación e Hipótesis

Integrantes del grupo:

Diego Sepúlveda Millán
Miguel Ángel Matas Rubio
Pedro Rodríguez Viñuales
David Rivera Concepción

1.Hipótesis.-

Los datos han sido tratados en busca de la facilitación a la tarea de responder las siguientes hipótesis:

1ª. La pandemia por la covid-19 provocó escasez de verduras y hortalizas en puntos clave de la pandemia.

2ª. Durante la pandemia los precios de las verduras y hortalizas se vieron afectados por las olas.

3ª. Los super alimentos han influido en la reducción de mortalidad por la covid-19.

4ª. En las comunidades donde es rica la ingesta de verduras y frutas hubo una menor tasa de mortalidad por covid.

2. Preproceso, transformación y tarjeta de datos.-

- *DatosDeConsumoAlimentario*:
    - Se han eliminado filas con todos los datos a 0, no aporta ninguna información.

	Año	Mes	CCAA	Producto	Volumen (miles de kg)	Valor (miles de €)	Precio medio kg	Penetración (%)	Consumo per capita	Gasto per capita
0	2018.0	Enero	Total Nacional	TOTAL PATATAS	108430.72	84640.08	0.78	79.40	2.38	1.85
1	2018.0	Enero	Total Nacional	PATATAS FRESCAS	79445.66	54688.29	0.69	68.46	1.74	1.20
2	2018.0	Enero	Total Nacional	PATATAS CONGELADAS	3999.90	4857.79	1.21	12.06	0.09	0.11
3	2018.0	Enero	Total Nacional	PATATAS PROCESADAS	4997.03	25094.00	5.02	45.94	0.11	0.55
4	2018.0	Enero	Total Nacional	T.HORTALIZAS FRESCAS	209957.24	376688.56	1.79	97.27	4.60	8.25
...	...	...	...	...	...	...	...	...	...	...
26508	2020.0	Junio	Valencia	PIÑA	917.49	1482.85	1.62	18.87	0.20	0.32
26509	2020.0	Junio	Valencia	OTRAS FRUTAS FRESCAS	2301.21	5908.75	2.57	50.99	0.49	1.27
26510	2020.0	Junio	Valencia	POMELO	202.67	460.37	2.27	5.07	0.04	0.10
26512	2020.0	Junio	Valencia	MANGO	193.14	507.45	2.63	9.15	0.04	0.11
26513	2020.0	Junio	Valencia	FRUTAS IV GAMA	1854.70	3947.11	2.13	27.41	0.40	0.85

25627 rows × 10 columns

```
Año                      float64
Mes                       object
CCAA                      object
Producto                  object
Volumen (miles de kg)    float64
Valor (miles de €)       float64
Precio medio kg          float64
Penetración (%)          float64
Consumo per capita       float64
Gasto per capita         float64
```

- *DatosMercaMadrid*:
    - Se han eliminado filas con todos los datos a 0
    - Se han eliminado las filas con volumen de producto 0
    - Se han eliminado las filas repetidas
    - Se han transformado los tipos de datos que eran números enteros de tipo object a int64
    - Se han cambiado el punto de los decimales de los float, por comas.
    - Se realizó un análisis del dataset a través de ProfileReport del PandasProfiling

YEAR	product	variedad	origen	Unidad	MONTH	price_mean	price_min	price_max		Volumen
2018	ACEITUNAS		ACEITUNAS		ALMERIA	kg 2	3,46	3,31	3,61	6700
2018	ACEITUNAS		ACEITUNAS		ALMERIA	kg 4	3,46	3,31	3,61	400
2018	ACEITUNAS		ACEITUNAS		ALMERIA	kg 5	3,46	3,31	3,61	260
2018	ACEITUNAS		ACEITUNAS		BADAJOZ	kg 4	3,46	3,31	3,61	24060
2019	ACEITUNAS		ACEITUNAS		BARCELONA	kg 2	3,46	3,31	3,61	1000
2020	ACEITUNAS		ACEITUNAS		BARCELONA	kg 7	3,46	3,31	3,61	2000
2018	ACEITUNAS		ACEITUNAS		CACERES	kg 2	3,46	3,31	3,61	700
2020	ACEITUNAS		ACEITUNAS		CACERES	kg 6	3,46	3,31	3,61	2000
2019	ACEITUNAS		ACEITUNAS		CADIZ	kg 8	3,46	3,31	3,61	28800
2019	ACEITUNAS		ACEITUNAS		CIUDAD REAL	kg 6	3,46	3,31	3,61	2000

```
product         object
variedad        object
origen          object
Unidad          object
familia         object
MONTH            int64
price_mean     float64
price_min      float64
price_max      float64
Volumen          int64
```

- *DatosCoranavirusCases*:
    - Se han eliminado las filas que no corresponden con España por lo tanto se han suprimido varias columnas y todo el dataset será tratado como total nacional, eliminando así todas las columnas que hacían referencia al país o territorio.

- Los valores “Nuevos fallecidos” y “Nuevos casos” han sido acumulados por meses.
- Se han unido las filas por mes, lo que significa que los datos de fallecimientos y contagios han sido sumados para poder agruparlos por meses y así tener un nexo de unión entre los demás data sets.
- Se ha normalizado la fecha mes:= int de dos dígitos y año:= int de 4 dígitos.
- Por último los casos acumulados durante 14 días atrás por cada 100.000 habitantes han sido normalizados con una media geométrica dando más peso a los datos de final de mes pero teniendo en cuenta todos los días para obtener una media aproximada de casos acumulados por cada mes.

3. Enriquecimiento de datos.-

3.1. Dataset1.-

El dataset 1 “*DatosDeConsumoAlimentario*” ha sido expandido gracias a tecnicas de web scraping en la pagina:
“https://www.mapa.gob.es/app/consumo-en-hogares/consulta11.asp”.
El rango de fechas del que disponemos ahora es desde 2013 hasta 2020, es decir se han añadido los datos desde 2013 hasta 2017. Y no solo se ha incrementado el rango de fechas, ahora disponemos de de datos de 27 categorías más de productos, aunque no nos centraremos en estos datos, de otros productos, pueden ser de interés una vez el estudio de las hipótesis esté mucho más avanzado, para poder contrastar conclusiones o reforzarlas.

Para crear el web scraper se ha usado la librería para Python3, Selenium. Esta librería permite capturar los elementos del navegador e interactuar con ellos, requerimiento indispensable para nuestro web scraper pues para acceder a los datos se debe rellenar un formulario indicando los datos que el cliente web desea visualizar. Una vez que se controla el formulario de acceso a los datos se pasa a tomar los elementos de la tabla que muestra la página web ordenadamente y darles formato CSV para poder usarlos al igual que los datos en crudo que se proporcionaron al inicio de esta asignatura.

Estos datos ampliados se usarán principalmente para determinar cuáles son los periodos normales en el consumo y no confundirlos con descensos provocados por la pandemia o situaciones que pudo desencadenar. Por ejemplo muchos de los alimentos que estamos tratando son de temporada, es decir se consumen durante ciertas temporadas del año pues el ciclo de cultivo y recolecta de muchos alimentos no permite que sea de otra manera, estos ciclos no los debemos malinterpretar durante los periodos de pandemia y señalarlos como causa directa pues estaríamos encontrado relaciones falsas.

3.2. Dataset5.-

Teniendo en cuenta los casos de COVID a nivel mundial, nos vimos con la necesidad de tener los datos españoles por Comunidad Autónoma. Para ello, después de una exhaustiva búsqueda en internet dimos con la página del portal estadístico. En ella encontramos una gran cantidad de información relativa a los casos de covid y fallecimiento por este a escala nacional. No solo tenemos los datos mensuales si no que tenemos que datos relativos a cada día:
https://portalestadistico.com/?pn=portalestadistico&pc=AAA00&idp=57&idpl=1348&idioma=
Con esta información y sumando todos los casos de todos los días y los fallecimientos, además de limpiar la base de datos con columnas que no se utilizan, nos quedamos con una amplia base de datos relativa a los casos confirmados y fallecimientos mensualmente. La tarjeta de datos queda de la siguiente manera:

	CCAA	Mes	CasosConfirmados	Fallecidos	Year
0	Andalucía	3.0	16903.0	755.0	2020.0
1	Andalucía	4.0	316451.0	25210.0	2020.0
2	Andalucía	5.0	423448.0	41461.0	2020.0
3	Andalucía	6.0	266148.0	29629.0	2020.0
4	Aragón	3.0	6569.0	386.0	2020.0

Mes	float64
CasosConfirmados	float64
Fallecidos	float64
Year	float64

3.3.Enriquecimiento clasificación productos.-

Se ha creado un pequeño web scraper para obtener los nombres de productos que son calificados como superalimentos con el objetivo de contestar la hipótesis número 3 centrándonos en comprobar si dichos alimentos han tenido algún tipo de relevancia.

4.Líneas de trabajo:

4.1.Línea de trabajo hipótesis 1.-

Comprobar con los datos del dataset 1, si la proporción de consumo per cápita y el gasto per cápita de los productos aumentó o bajó en igual proporción. Si por el contrario por ejemplo aumentó en proporción más el gasto per cápita que el consumo per cápita entonces significaría que hubo más escasez, en casos en donde esto se acentuará mucho se podría hablar de que sí hubo escasez en ese momento y se comprobaría con que momento de la pandemia coincidió. Con esto se podría analizar la relación entre los casos covid y la escasez y poder crear un modelo capaz

de predecir posible escasez de ciertos productos en una pandemia con características similares.

4.2.Línea de trabajo hipótesis 2.-

Primero se va a contabilizar los contagios por meses sumando los contagios diarios del dataset 5 "*Coronavirus cases*". Después se van a buscar reglas de asociación con los diferentes productos del dataset 1 "*DatosDeConsumoAlimentario*".

4.3.Línea de trabajo hipótesis 3.-

Con el total nacional de consumo, se comprobará si hay una relación directa entre la penetración de ciertos alimentos denominados superalimentos y compararlos con los datos de contagios y fallecimientos, teniendo en cuenta los descensos por fin de oleada o por estacionalidad.

4.4.Línea de trabajo hipótesis 4.-

Averiguar si el consumo de verduras y de frutas por comunidad autónoma, tiene relación con la tasa de mortalidad por coronavirus. Para ello se realizará un estudio de los casos confirmados de COVID junto con sus muertes. La medida se hará mensualmente y se calculará la tasa de mortalidad con una simple división entre las columnas de datos confirmados y fallecimientos.

5.Entornos de trabajo.-

Estamos trabajando tanto en lenguaje python como en cuadernos de Colab. Todo ello se va registrando a medida que se sube al github del grupo:

https://github.com/DiegoSM1998/MineriaMultiagentesAGRO