

UNIVERSITYHACK 2021®
DATAATHON



The Data Masters

Autores:

- Elene Astondo Gutierrez
- Nagore Bermeosolo Saitua
- Unai Torrecilla López

Enlace a la aplicación:

- <https://datamasters.shinyapps.io/TheDataMasters/>

Índice de contenido

1	Objetivos.....	2
2	Stack Tecnológico utilizado.....	3
3	Fase 1: preprocesamiento de los datos.....	3
4	Fase 2: procesamiento de los datos.....	3
5	Fase 3. Implementación de la aplicación visual interactiva.....	3
6	Estructura de la interfaz.....	4
7	Contenido de la aplicación.....	5
8	Información adicional.....	6
9	Líneas futuras.....	6

Índice de figuras

FIGURA 1: Arquitectura general del proceso seguido para la implementación de la aplicación de analítica visual.....	2
FIGURA 2: Diferentes secciones del layout.....	4

Para la resolución del reto planteado por Cajamar, se ha creado una aplicación interactiva mediante la cual se pueden extraer diversas conclusiones sobre el impacto que ha tenido la *covid-19* en el mercado de frutas y hortalizas español.

La **principal motivación** del equipo ha sido conseguir que el usuario sea **partícipe** de este proceso. De esta manera, se le brinda la posibilidad de extraer conocimiento de manera clara y sencilla gracias a la herramienta de analítica visual¹ interactiva implementada.

Del mismo modo, con el objetivo en mente de enriquecer las conclusiones obtenidas y de reducir la carga cognitiva a la hora de interpretar las conclusiones obtenidas, se han añadido bloques explicativos y orientativos facilitando al usuario la interpretación de los datos. A su vez, se ha trabajado minuciosamente la usabilidad de la aplicación, la experiencia de usuario y el diseño visual de la misma, factores clave en la visualización de información², obteniendo como resultado una aplicación *user-friendly*.

En la Figura1 se puede observar el proceso que se ha seguido desde la ingesta de los datos hasta la implementación de la aplicación:

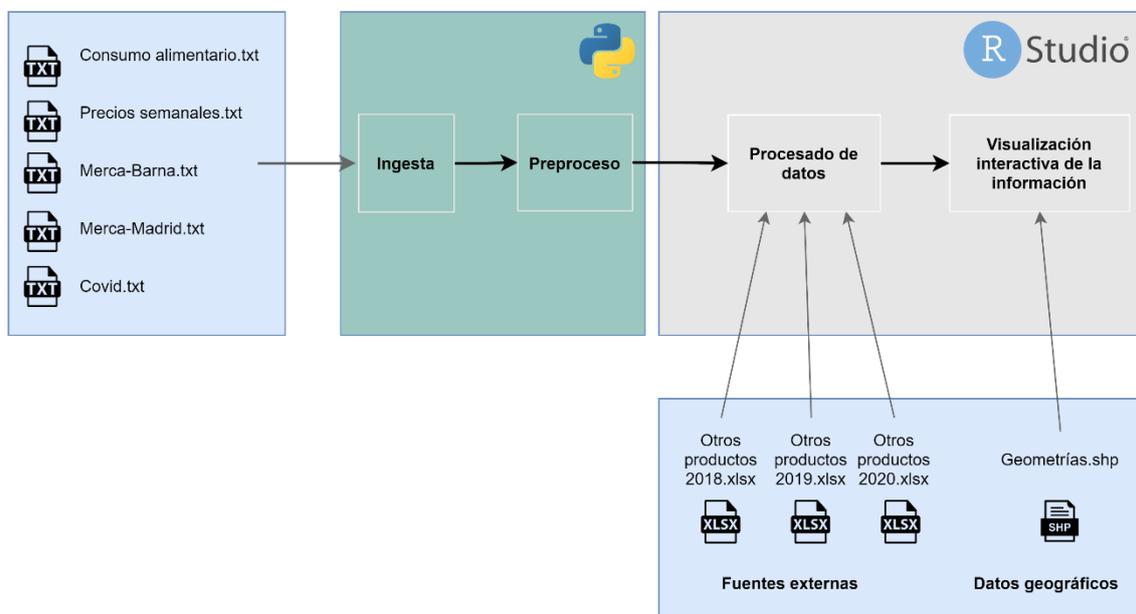


FIGURA 1: Arquitectura general del proceso seguido para la implementación de la aplicación de analítica visual

1 Objetivos

La aplicación implementada tiene como objetivo principal dar respuesta a las siguientes preguntas:

1. ¿Cuál es el estado del mercado de frutas y hortalizas español?
2. ¿Cuánto ha afectado el virus a este mercado?
3. ¿La alimentación ha sido más saludable en el periodo de pandemia?

¹ Keim, Daniel, et al. "Visual analytics: Definition, process, and challenges." *Information visualization*. Springer, Berlin, Heidelberg, 2008. 154-175.

² Cawthon, Nick, and Andrew Vande Moere. "The effect of aesthetic on the usability of data visualization." *2007 11th International Conference Information Visualization (IV'07)*. IEEE, 2007.

2 Stack Tecnológico utilizado

Como lenguajes de programación se han utilizado Python para la fase de preprocesamiento y R para el desarrollo de la aplicación interactiva y visualización de información. A su vez, se han utilizado las siguientes librerías:

Librerías utilizadas	
Python	RStudio
Pandas	Shinydashboard
Numpy	Ggplot2
Statsmodel (para el modelaje)	Dplyr

3 Fase 1: preprocesamiento de los datos

Previo a la creación de las visualizaciones y la aplicación, ha sido necesario realizar una limpieza general de los datos proporcionados. Esta fase ha consistido principalmente en las siguientes tareas:

- Tratamiento de valores nulos: comprobación e imputación.
- Clasificación de las variables del dataset.
- Tratamiento de variables textuales: esta tarea ha consistido en adecuar todas las entradas de texto de las que se disponían al lenguaje castellano.
- Eliminación de información redundante.

El output generado en esta fase de preproceso, servirá como base desde la cual se trabajará en etapas posteriores del proyecto.

4 Fase 2: procesamiento de los datos

Además de la fase de preproceso, de cara a generar los modelos de visualización correspondientes, se ha llevado a cabo un tratamiento adicional de los datos para realizar las visualizaciones correspondientes. Las tareas realizadas en esta fase han sido:

- Agregación de los datos por promedios o sumas.
- Creación de nuevas variables.
- Unificación de todos los campos textuales para que la aplicación tenga coherencia.

5 Fase 3. Implementación de la aplicación visual interactiva

Como se puede observar en la figura 2, la aplicación visual desarrollada se encuentra dividida en cuatro secciones:

- Menú: Permitiendo al usuario la navegación entre las diferentes interfaces
- Filtros: Zona interactiva para seleccionar las variables a inspeccionar
- Objetos visuales: Área donde se incorporan los modelos de visualización
- Conclusiones: Área donde se añaden las conclusiones obtenidas.



FIGURA 2: Diferentes secciones del layout

6 Estructura de la interfaz

Se ha desarrollado una aplicación visual e interactiva mediante el uso de Shiny y RStudio. La interfaz de la aplicación consta de 3 páginas y cada una de ellas da respuesta a los objetivos planteados previamente (página 2). Estas a su vez contienen diferentes subpáginas que se centran en objetivos más específicos:

1. Página 1: Agroanálisis

- 1.1 **Análisis general:** dilucidar el comportamiento de los productos en diferentes comunidades autónomas en distintas épocas del año.
- 2.1 **Comparación productos:** comparar el comportamiento de diferentes productos en diferentes comunidades autónomas.

2. Página 2: Impacto Covid

- 1.1 **Impacto general:** cuantificar el impacto del virus en el mercado de frutas y hortalizas.
- 2.1 **Análisis temporadas:** cuantificar el impacto del virus en las temporadas de los productos.
- 3.1 **Top productos:** analizar los cambios de tendencia en el consumo de productos derivados de la pandemia.
- 4.1 **Importaciones y exportaciones:** observar el impacto del virus en las exportaciones e importaciones de España con respecto a otros países.

3. Página 3: Impacto salud

- 1.1 **Familias de productos:** concluir mediante el consumo de diferentes familias de productos si la alimentación ha sido en general más saludable durante el periodo de pandemia o no.

El objetivo del reto a resolver ha sido el de visualizar el impacto que la pandemia ha tenido en el mercado de frutas y hortalizas, entonces ¿por qué se ha usado esta estructura? Se pretende que el usuario a través de una interfaz visual interactiva tenga la posibilidad de conocer el estado de este mercado en la situación previa a la pandemia, para después poder visualizar el impacto que esta ha tenido. Finalmente se contrasta la hipótesis, tan escuchada en los medios de comunicación, de si en este periodo el consumo alimentario ha sido realmente más saludable o no.

7 Contenido de la aplicación

7.1 Agroanálisis

7.1.1 Análisis general

Se pretende realizar un análisis general del precio o el consumo per cápita de cada uno de los productos. Para ello, se ha creado un mapa geográfico de calor. Una conclusión que se puede obtener de esta página es que, en valencia, el consumo per cápita de las alcachofas duplica la media nacional, esto puede deberse a que el plato típico 'Alcachofas de Benicarló' tiene como ingrediente principal este producto.

7.1.2 Comparación productos

El objetivo principal de esta página es comparar las características de diferentes productos en varias comunidades. Para ello se ha añadido un gráfico que muestra la información nacional. A este le acompañan los gráficos correspondientes a las comunidades autónomas que se seleccionen. El motivo de esta estructura es que, en los medios de comunicación, se muestran las visualizaciones correspondientes a la covid-19 de esta manera.

En esta página, por ejemplo, se puede apreciar que el precio de los ajos es mucho más estable, y normalmente más elevado, en la comunidad de Asturias que en Andalucía o la media nacional.

7.2 Covid

7.2.1 Impacto general

Esta página tiene como objetivo cuantificar el impacto que ha tenido el virus en el mercado de frutas y hortalizas español. Para ello, se han realizado predicciones con modelos de la familia ARIMA para observar lo que habría sucedido de no ocurrir la pandemia. Mencionar, que se ha realizado una predicción para cada producto y comunidad autónoma.

En esta página, primeramente, se muestra un gráfico en el que se muestra una línea de los datos reales y otra línea discontinua que hace referencia a la predicción realizada. Por otra parte, se visualiza mediante un mapa geográfico la diferencia entre el valor real y las predicciones obtenidas por el modelo. De este modo se puede concluir si el impacto del virus sobre este mercado ha sido positivo o negativo.

Por ejemplo, se puede observar como la pandemia ha supuesto que el precio de las frutas de cuarta gama disminuya 0,45 €/kg en Murcia.

7.2.2 Análisis temporadas

Las frutas y las hortalizas se caracterizan normalmente por ser productos marcadamente estacionales. Por este motivo, esta página cuantifica el impacto del virus en las temporadas de cada producto mediante dos mapas geográficos de calor, uno con los datos previos a la pandemia y otro con los datos de la pandemia.

En esta pantalla se puede ver por ejemplo que la temporada de los espárragos ha sido bastante perjudicada en Andalucía habiendo supuesto casi un 25% de decremento en los ingresos derivados de este producto.

7.2.3 Top productos

Para poder observar si las tendencias de consumo de productos han cambiado como consecuencia de la pandemia, se muestran dos gráficos de araña en los que se visualizan los productos más consumidos durante la pandemia y en el periodo previo a ella. Como conclusión obtenida, en esta esta página se puede observar que, si bien ha habido cambios en las tendencias en algunas ocasiones, por lo general se han mantenido.

7.2.4 Importaciones y exportaciones

Esta página cuenta con la información referente al comercio exterior del mercado de frutas y hortalizas español. Se puede seleccionar el país cuyas exportaciones/importaciones se quieren analizar. Se muestra un gráfico de líneas con ambas actividades comerciales para poder compararlas. Por lo general, se puede comprobar que se ha importado y exportado menos.

7.3 **Salud**

7.3.1 Familia de productos

Durante la pandemia se ha escuchado en numerosas ocasiones que el consumo alimentario ha sido más saludable. Para poder contrastar esta hipótesis, se compara el consumo de frutas y hortalizas con otras familias de productos.

Si bien el consumo de las frutas y hortalizas ha aumentado durante y tras la pandemia, esta subida también se ha dado en otro tipo de producto como puede ser la leche, los huevos o la carne. Ahora bien, parece ser que el aumento en las frutas y hortalizas es superior.

8 Información adicional

Todo el código fuente del proceso/desarrollo y los pasos necesarios para desplegar la aplicación se pueden encontrar en la página de [GitHub](#). Para llevar a cabo esta tarea de manera sencilla se explican las instrucciones en el archivo README.md.

9 Líneas futuras

Como líneas futuras se plantean la mejora de la usabilidad y experiencia de usuario, el incremento del conjunto de datos de fuentes externas para optimizar los modelos implementados y como último, añadir nuevos modelos de visualización de información para mejorar la eficacia de la aplicación.