**Resultados**

Los resultados que se presentan a continuación en esta investigación fueron obtenidos siguiendo la metodología CRIPS-DM, en cual indica que como primera fase es el entendimiento del negocio, Pagar es Fácil es un eje de negocios digitales que permite a pequeños y medianos empresarios exhibir y vender productos o servicios, así como transaccionar con tarjetas de crédito y billetera virtual, transferir saldos a cuentas bancarias y pagar servicios básicos. Pagar es Fácil ofrece múltiples servicios, pero la que se utilizó en esta investigación fue el servicio de MarketPlace. Pagar es Fácil aloja muchos comercios los cuales día a día suben sus productos a la plataforma e igualmente se transacciona con las compras realizadas por las personas.

Regularmente la plataforma recibe ingresos económicos por comisiones de cada venta, siempre y cuando el vendedor entregue correctamente el producto. En base a lo mencionado anteriormente, Pagar es Fácil requiere conocer una predicción de ganancias de las ventas que se podrían generar en futuros años, conocer cuáles son sus compradores más recurrentes, las categorías y productos más solicitados, y sus vendedores más confiables.

Con respecto a la segunda fase, el entendimiento de los datos, EL dataset fue obtenido de las diferentes bases de datos proporcionadas por la plataforma, las cuales son las ventas conformadas por los clientes, vendedores, productos y categorías, a continuación, se detalla cada una ellas:

Cliente: Nombres y apellidos de los clientes.

Nombre\_producto: el nombre del producto.

Name\_comercio: el nombre del vendedor o mejor conocido como el comercio.

Precio\_producto: precio del producto proporcionado por el vendedor.

Precio\_producto\_venta: precio del producto sumado con la comisión de Pagar es Fácil.

Ganancias: es la resta de precio\_producto\_venta con precio\_producto obteniendo así las ganancias por producto por venta.

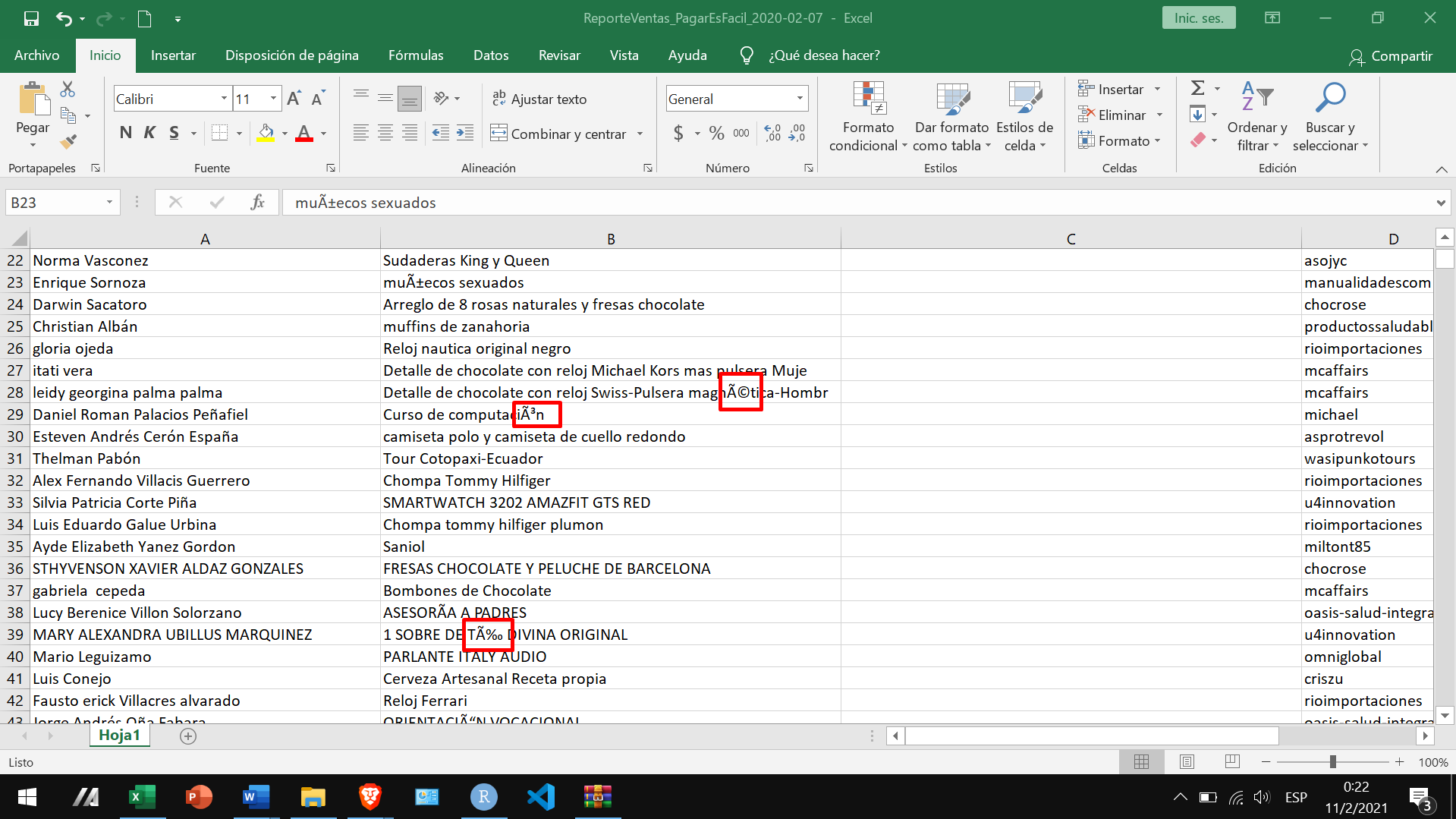
Category: las categorías de los productos.

Fecha: fecha de venta.

Venta\_concretada: variable utilizada para conocer si un vendedor entrego el producto al cliente.

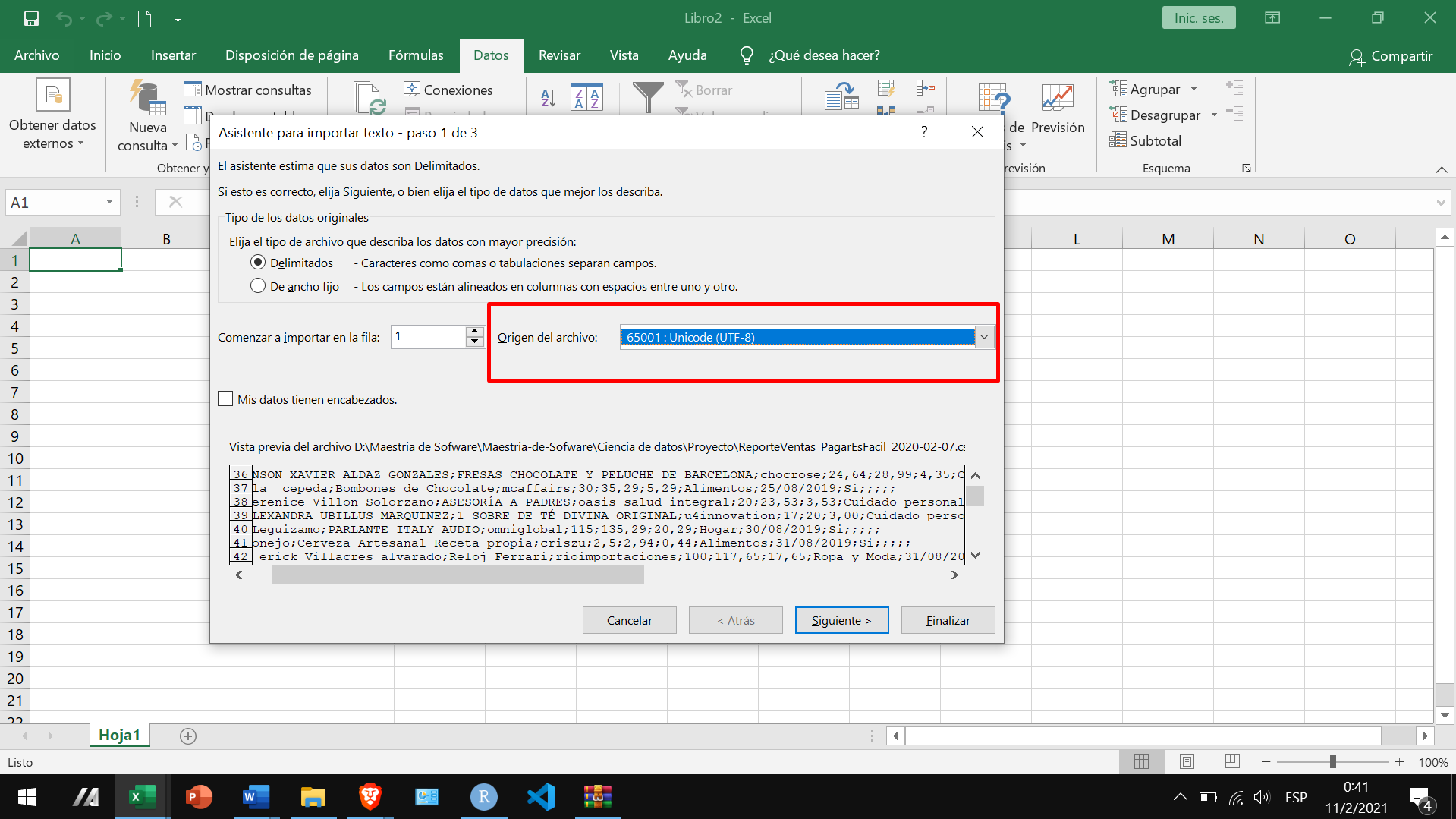
Durante la tercera fase, los datos pasaron por un pre procesamiento, las cuales se indican a continuación, lo primero que se hizo con el dataset fue corregir aquellos campos con caracteres que están mal codificados.

*Estado inicial de la data en Excel*



**Fuente:** Elaboración Propia

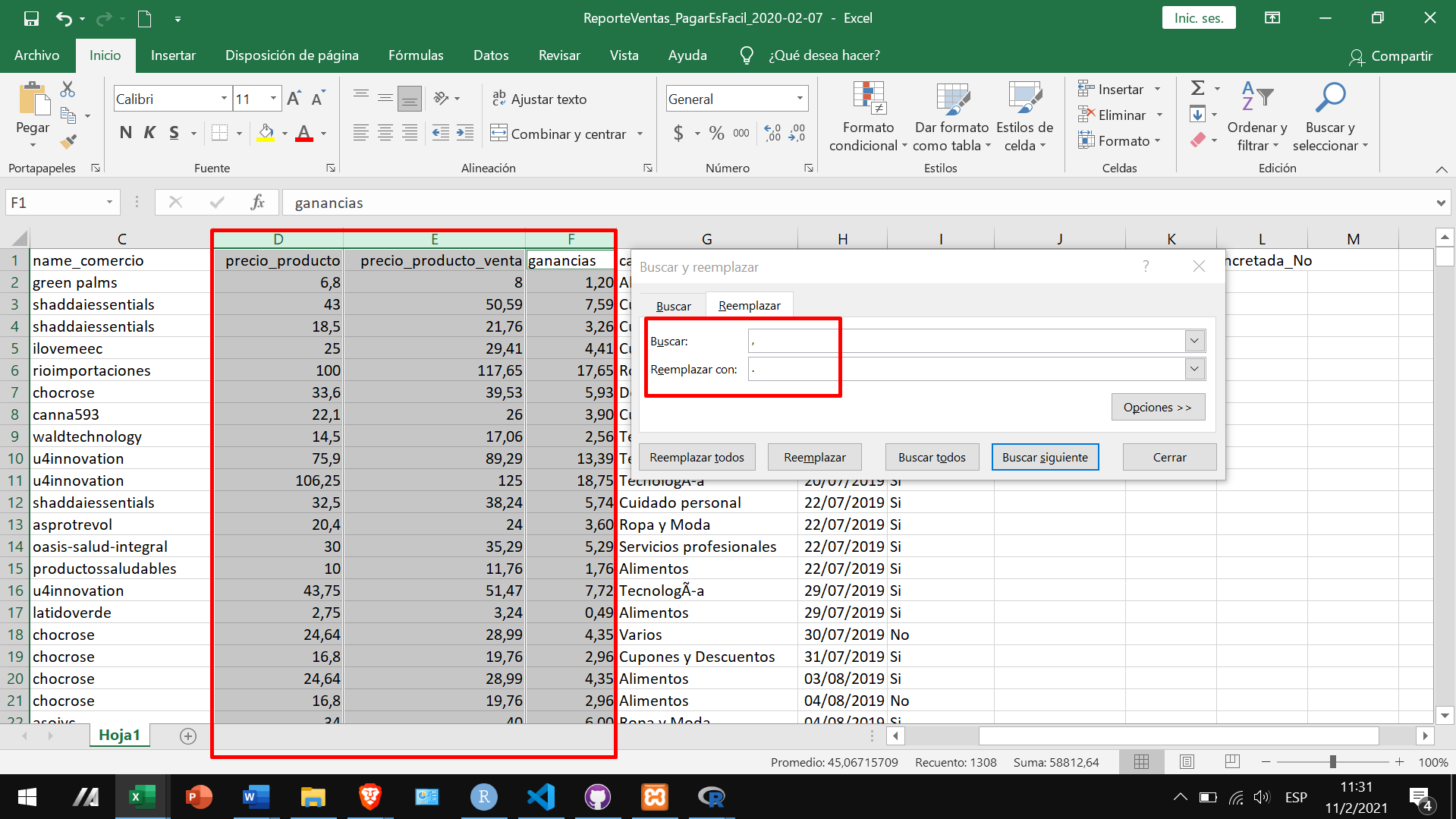
*Proceso de corrección de datos en Excel*



**Fuente:** Elaboración Propia

Se cambiaron las comas que poseen los campos numéricos de precio\_producto, precio\_producto\_venta y ganancias por el punto, para que posteriormente la herramienta R nos los reconozca como variable numérica y no una cadena de texto.

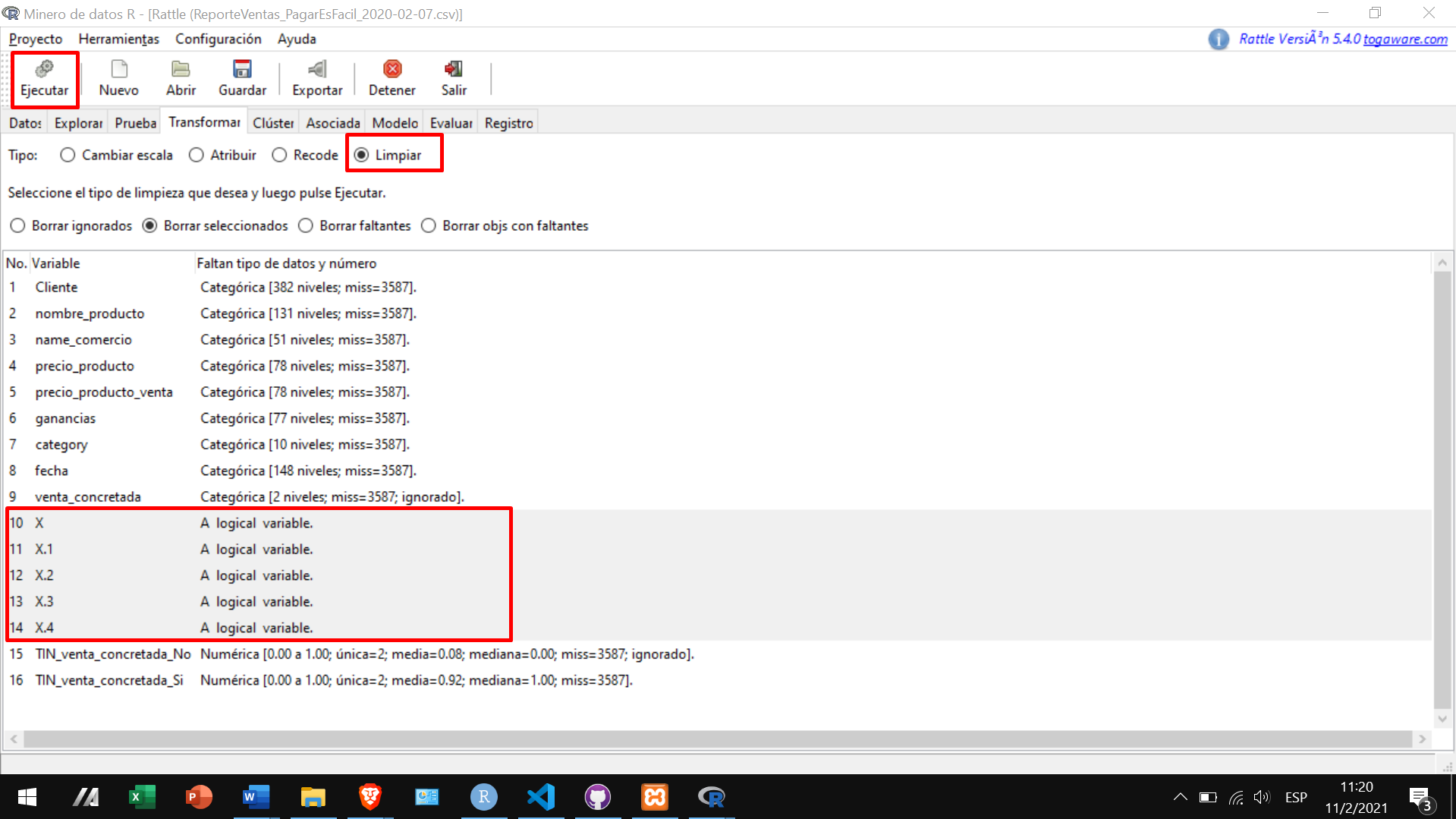
*Cambio del punto decimal al punto en valores numéricos en Excel*



**Fuente:** Elaboración Propia

Se borrarón aquellos campos adicionales que se crearon al transformar el archivo .xlms a .csv, para esto se utilizó la librería Rattle.

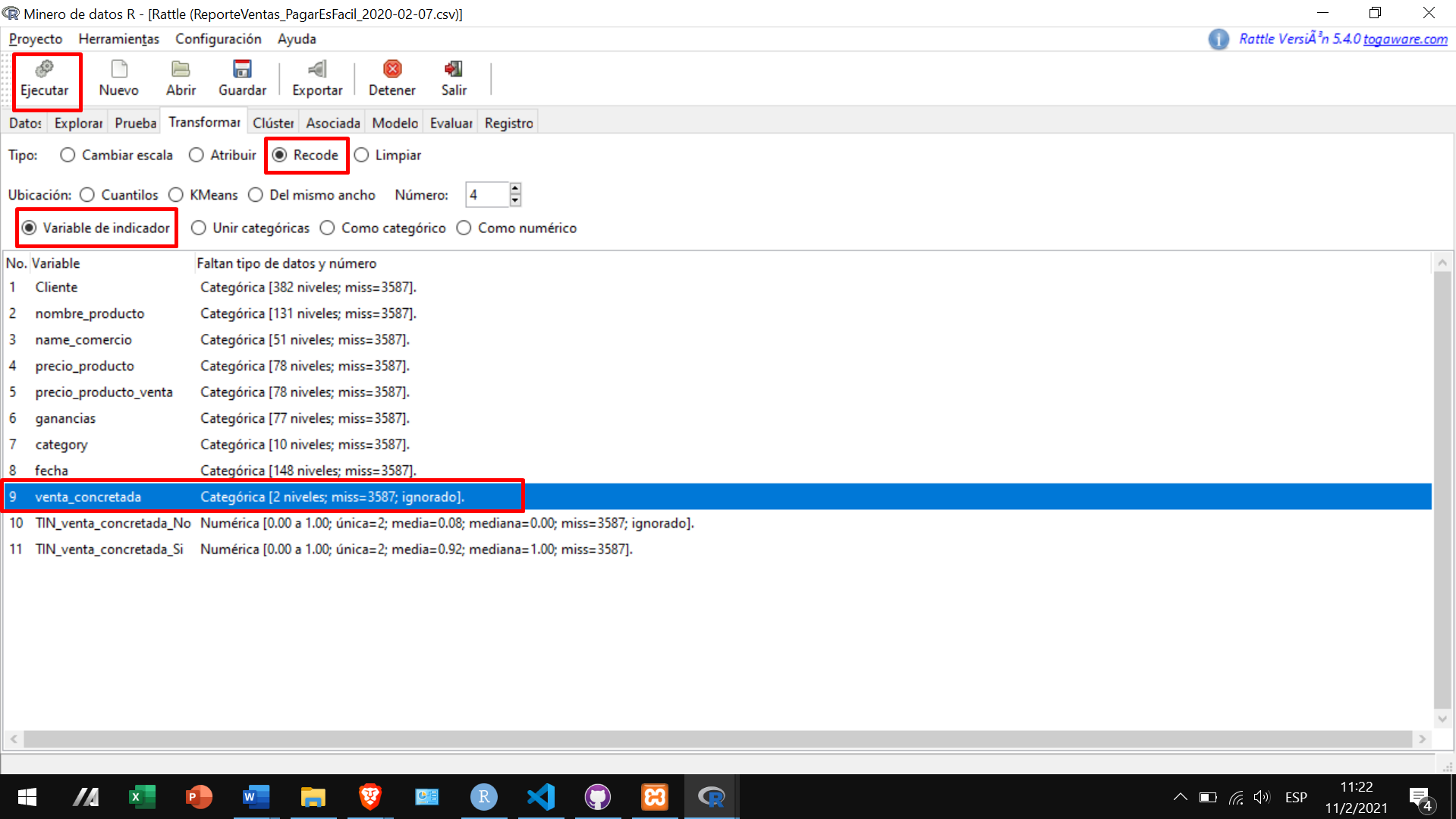
*Eliminación de campos adicionales de la data con Rattle.*



**Fuente:** Elaboración Propia

Se utilizó la opción RECODE de la librería Rattle, para crear variables indicadoras a la columna venta\_concretada que poseen registros con valores de SI/NO y posteriormente utilizarlas para un posible análisis.

*Creación de variable indicadora venta\_concretada con Rattle.*



**Fuente:** Elaboración Propia

A lo que respecta con la fase de modelamiento, se utilizó la nube de palabras para determinar cuáles son los productos más comprados en la tienda y también se aplicó el modelo ARIMA en las series temporales para la predicción de ganancias de las futuras ventas en Pagar es Fácil.

**Nubes de palabras**

En la siguiente imagen se detalla el resultado de esta técnica de minería de datos, en la cual se analizaron todos los productos vendidos desde el junio del 2019 hasta febrero del 2021, con una cantidad de frecuencia de por ejemplo 7 y un total de 89 palabras, dando como resultado que las palabras chocolate, aceite, fresas y computador fueron las más buscadas y por ende vendidas por la tienda.

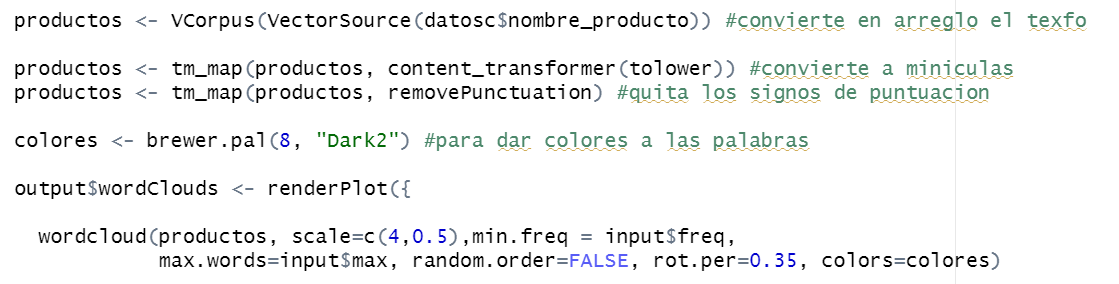
*Empleo de la nube de palabras en R.*



***Fuente:*** *Elaboración Propia*

El código utilizado para la creación de la nube de palabras fue el siguiente:

*Código fuente de la creación de nubes de palabras en R*

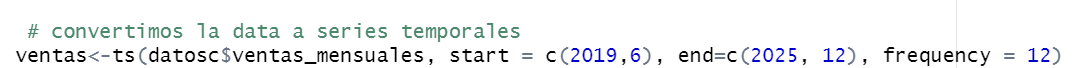


***Fuente:*** *Elaboración Propia*

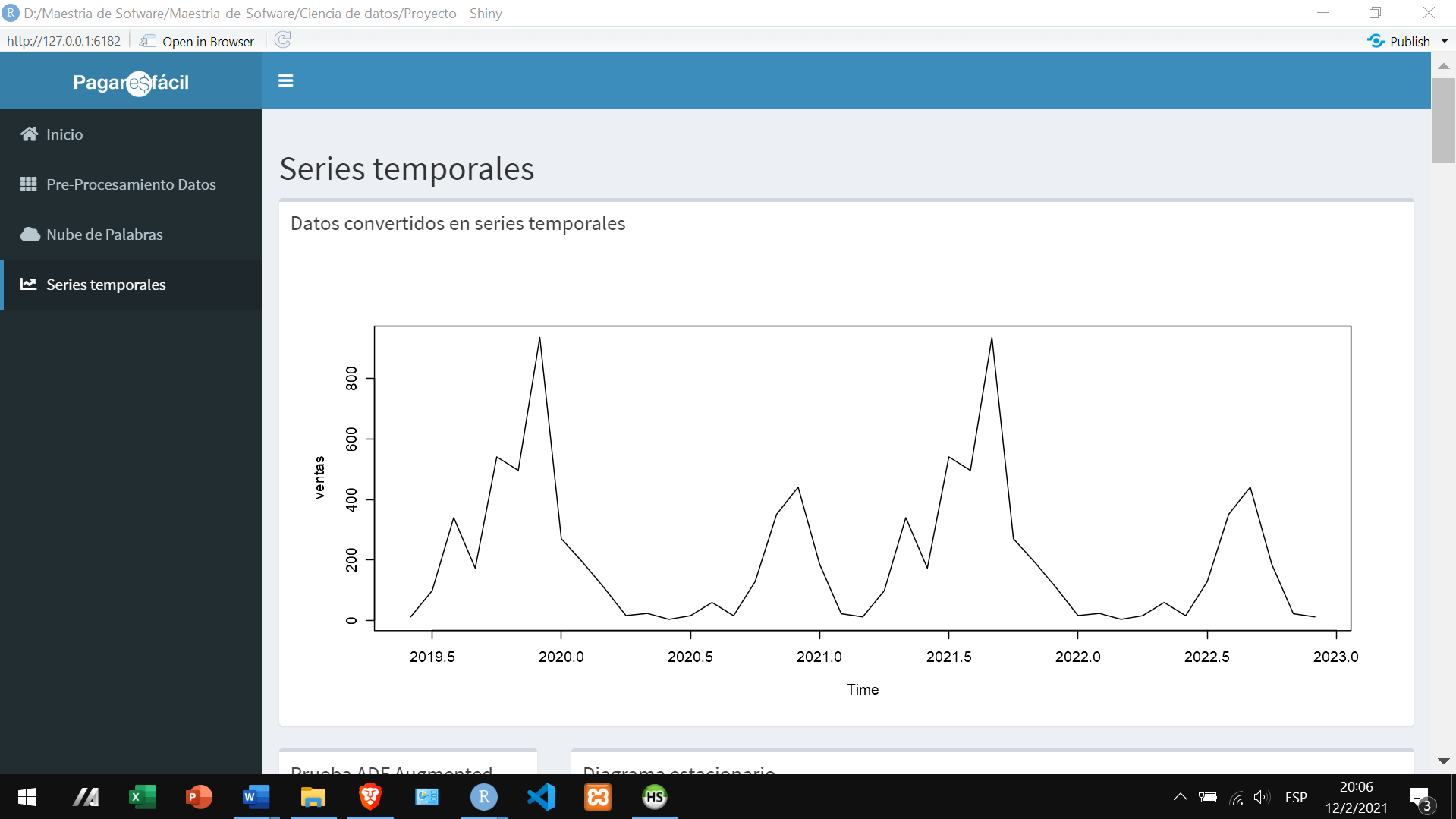
**Series temporales**

El primer paso para trabajar con las series temporales es convertir los datos en una serie temporal, se utilizó el siguiente código en donde se detalla el mes y año de inicio de la primera venta y el mes y año final en la que se quiere hacer

*Conversión de la data en serie temporales.*

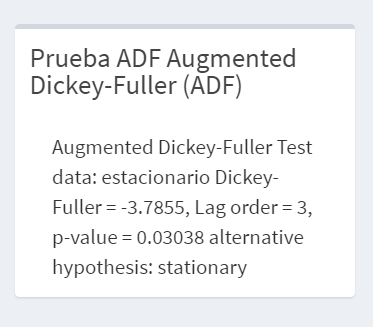


***Fuente:*** *Elaboración Propia*

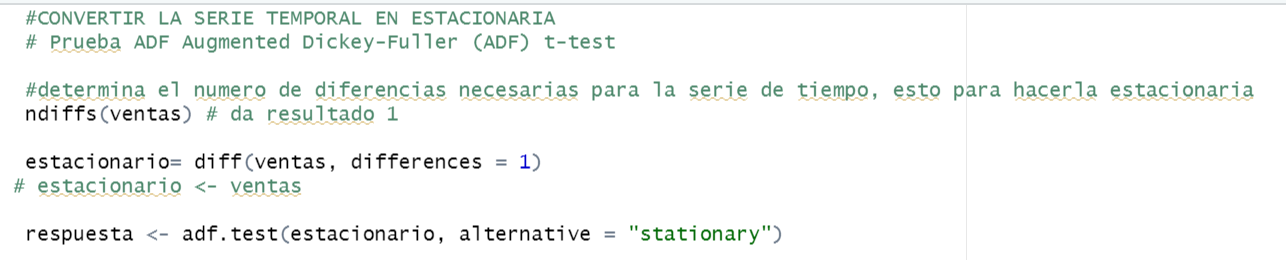


***Fuente:*** *Elaboración Propia*

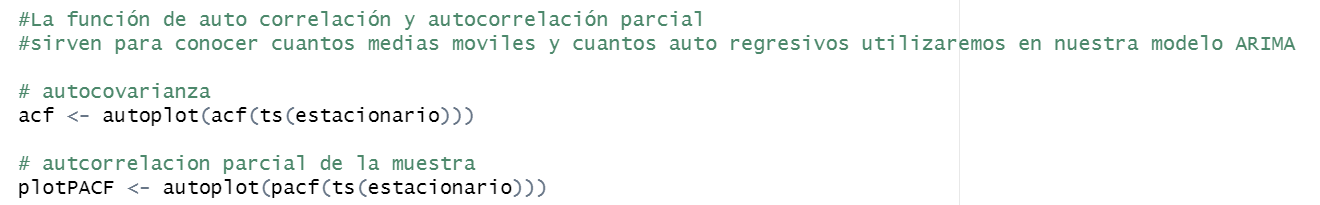
Para que se pueda aplicar el modelo Arima, un requisito es que la serie temporal debe ser estacionaria, una serie es estacionaria cuando los valores están cercas de una misma media, entonces para comprobar si la serie temporal es estacionaria se aplicó el test de Dickey Fuller el cual indica que la el p-value debe ser menor a 0.05, dando como resultado 0.03 lo cual indica que nuestra serie es estacionaria.



El código utilizado fue el siguiente:

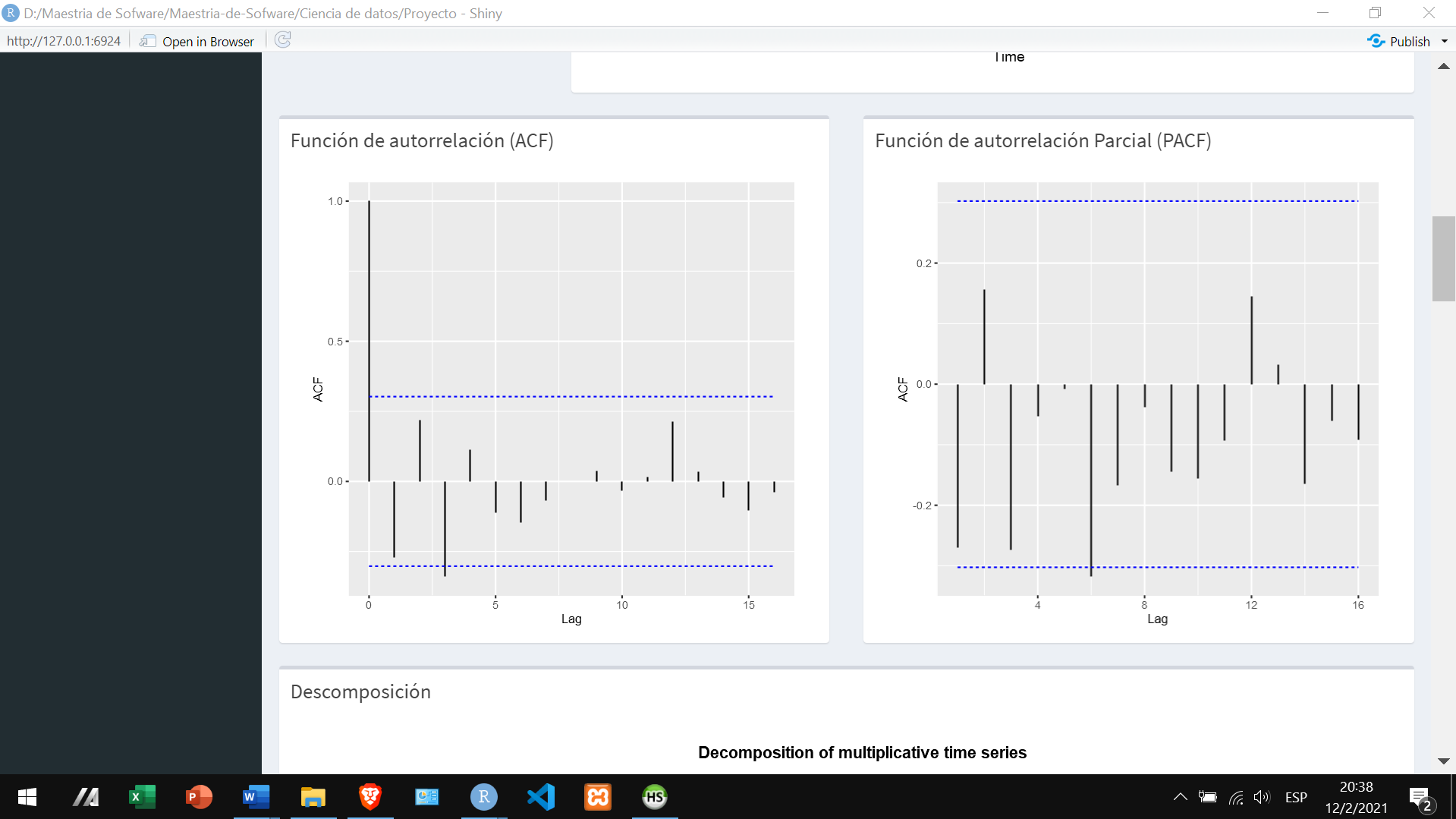


Luego se obtuvieron los valores de las funciones de auto correlación (ACF) y la función de auto correlación parcial (PACF), estas dos funciones nos sirven para saber cuántas medias móviles y cuantos autoregresivos vamos a utilizar en nuestro modelo ARIMA entonces para eso se utiliza el siguiente comando:



Analizando los resultados:

El ACF nos indica el número de medias móviles y el PACF nos indica el número de autoregresivos y en ambos nos aparecen los rezagos de la serie, del cual para el ACF el rezago es 3, y en PACF es 6, estos dos valores nos servirán posteriormente para la aplicación de ARIMA.



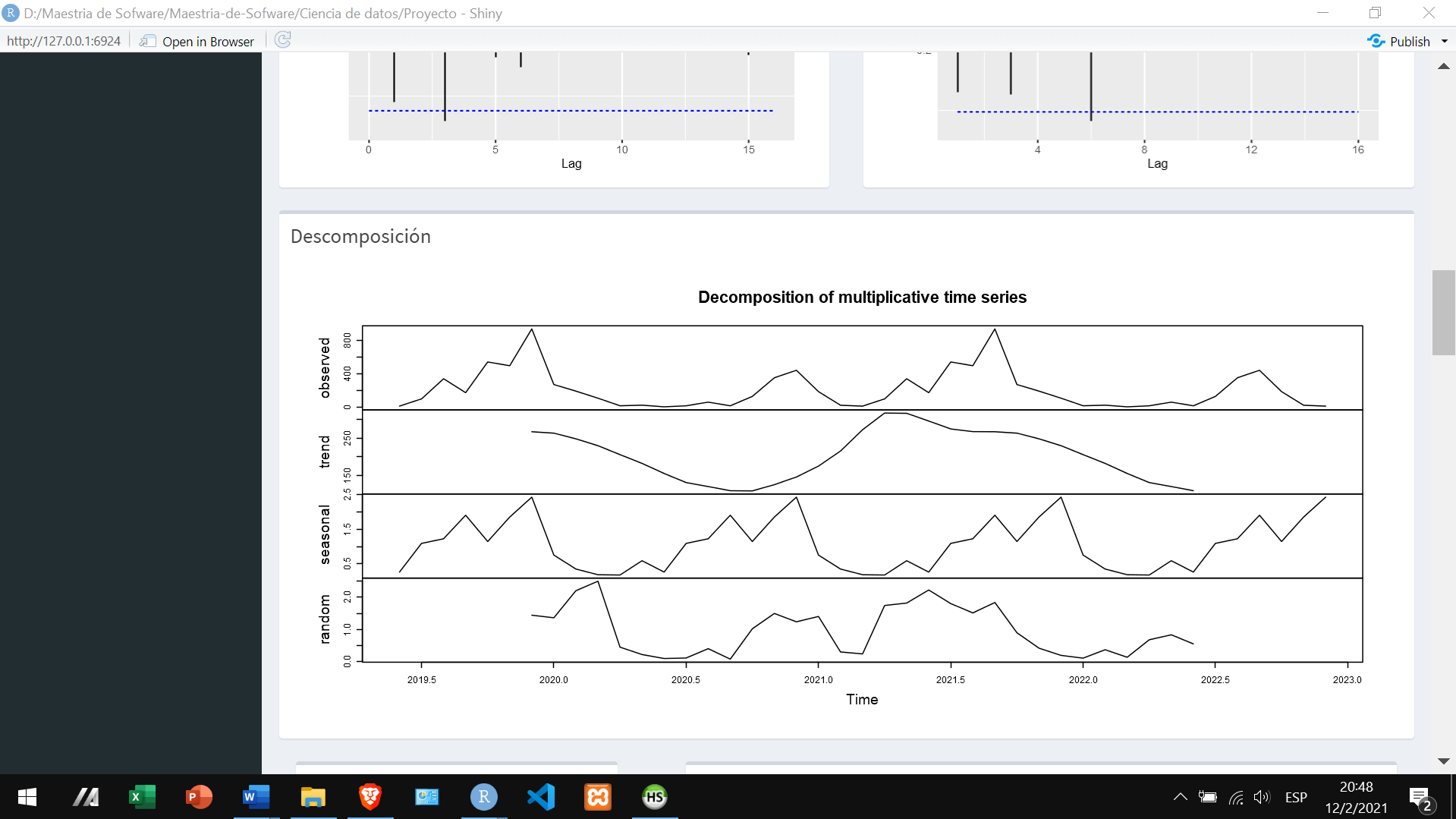
Hasta el momento tenemos de resultado:

ACF= 3

PACF= 6

Diferencias=1

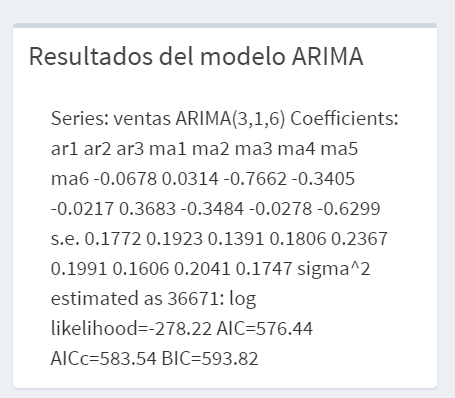
Se procedió a descomponer la serie para analizar más a fondo la tendencia, estacionalidad, componente irregular y los datos observados.



Para aplicar el modelo ARIMA, se utilizó los valores previamente obtenidos, los cuales son ACF, diferencias y PACF



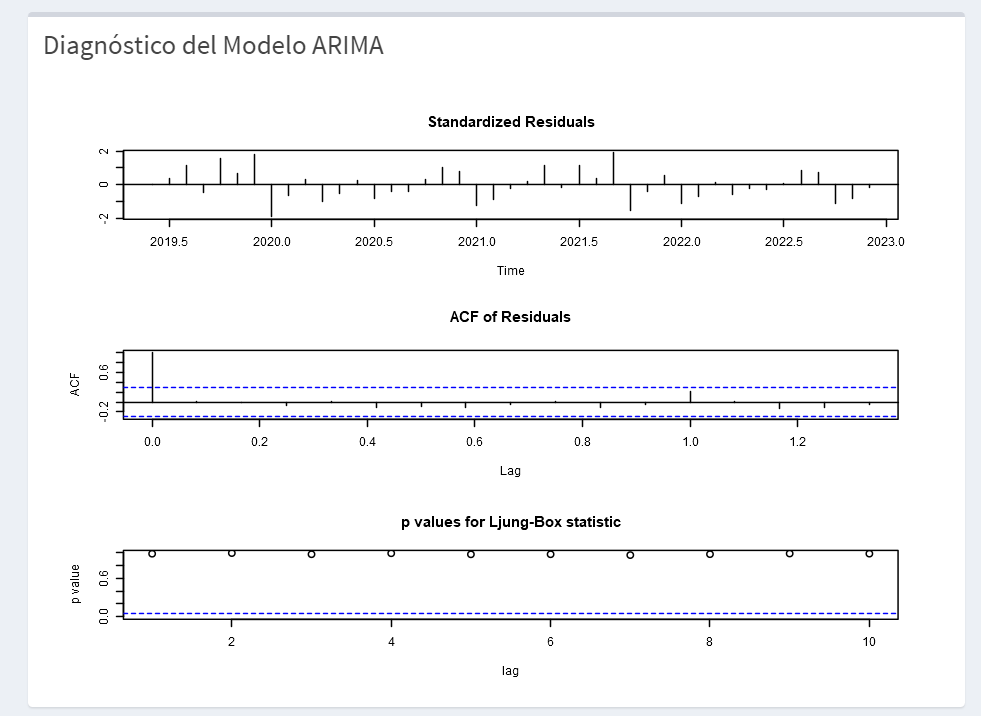
Obteniendo el siguiente resultado, en el cual nos indica los coeficientes de autoregresivos, los coeficientes de la media móvil y los errores estándar.



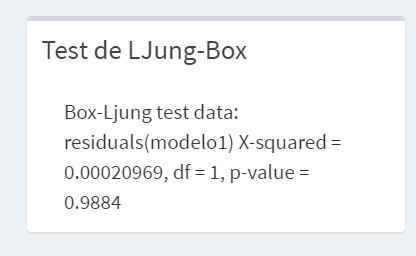
Se puede realizar un diagnóstico para saber si el modelo es bueno, se utiliza el siguiente comando.



La siguiente gráfica muestra un diagnóstico del modelo que utilizamos, en la primera gráfico se muestran los errores estandarizados y estos deben parecerse mucho al ruido blanco, pero para eso se hará uso de los valores p (p-values) que entrega la tercera gráfica, el estadístico de Ljung Box, los cuales nos indica si hay o no ruidos blanco, y como se aprecia en la gráfica #3 los valores son mayores a 0.5 lo que indica según el test de Ljung que si es mayor a 0.5 entonces hay ruido blanco y nuestro modelo se ajusta bien.

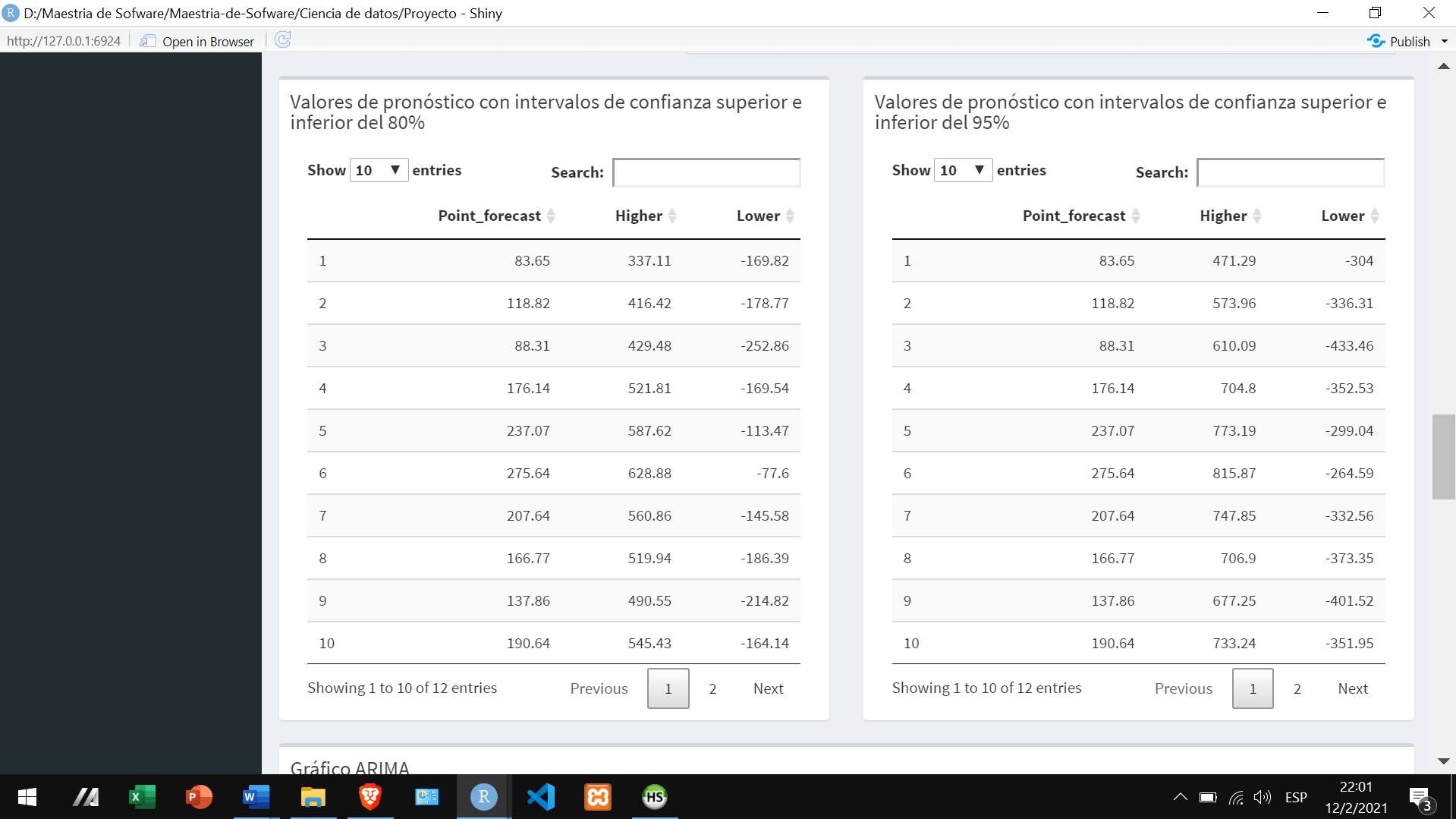


Para confirmar lo anteriormente dicho, se ejecutó el test de Ljung dando como resultado un p-value de 0.98 siendo este mayor a 0.5.



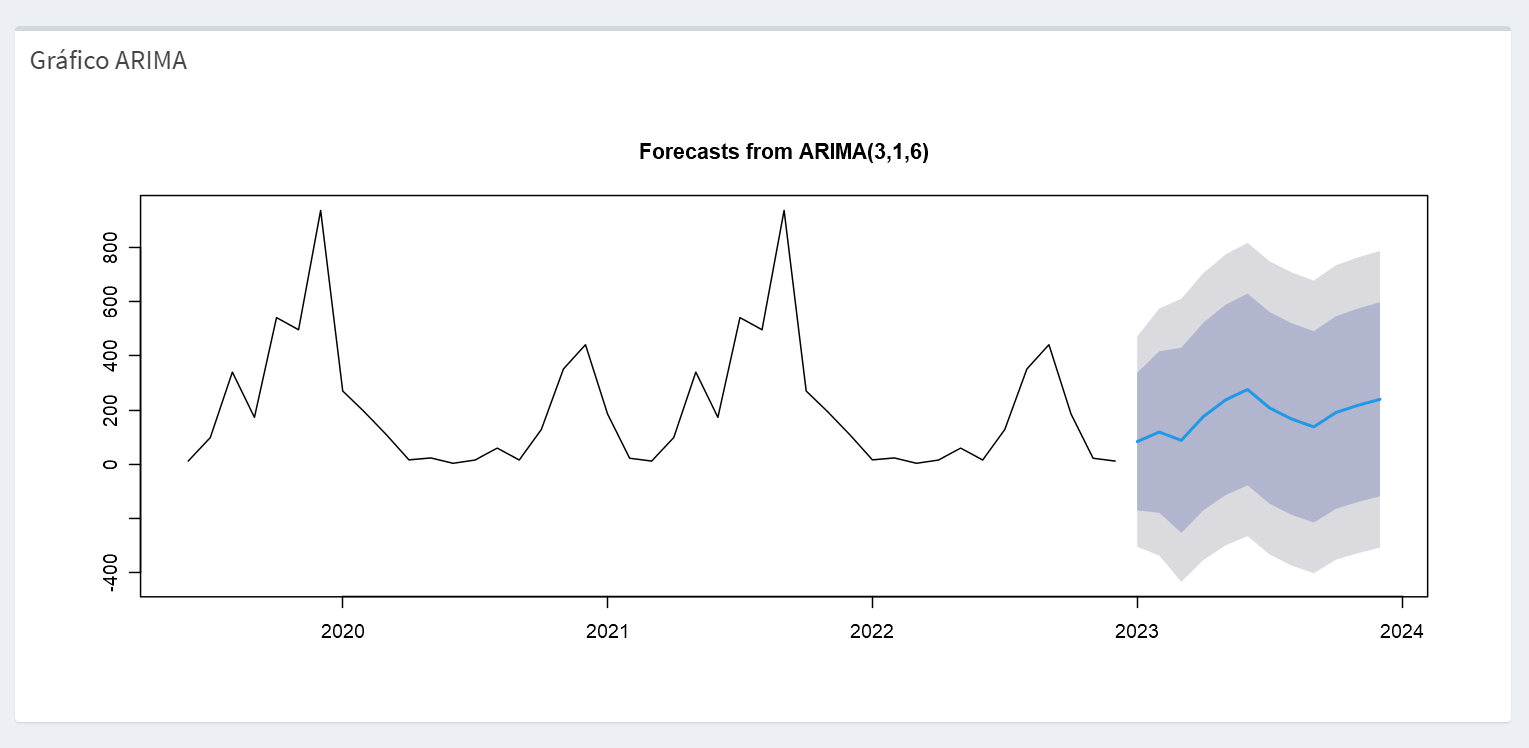
**Evaluación**

Para dar una predicción o pronostico de la cantidad de ganancias de ventas dentro de los próximos 12 meses de Pagar es Fácil hay que analizar los limites inferiores y superiores con el 80% y 95% de confianza.

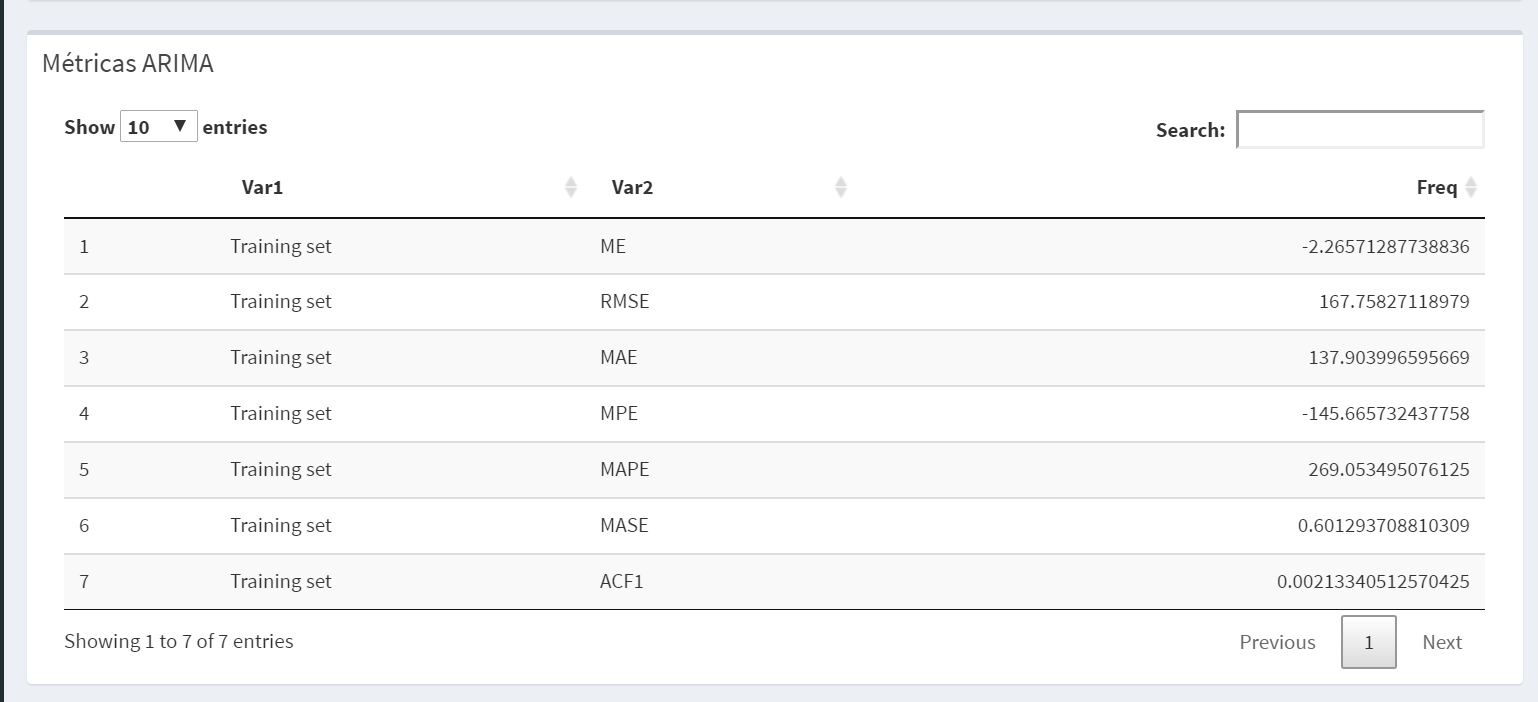


La interpretación del resultado indica por ejemplo, para el mes de enero del 2023, Pagar es Fácil tendrá un ingreso de $83.65 tanto a nivel de 80% y 95% de confianza, siendo su valor más alto $337.11 y su valor más bajo $-169.82 para el nivel de confianza de 80% y para el 95% su valor más alto es de $471.29 y su nivel más bajo $-304.

La siguiente gráfica muestra el resultado de la tendencia de las ventas a futuro.

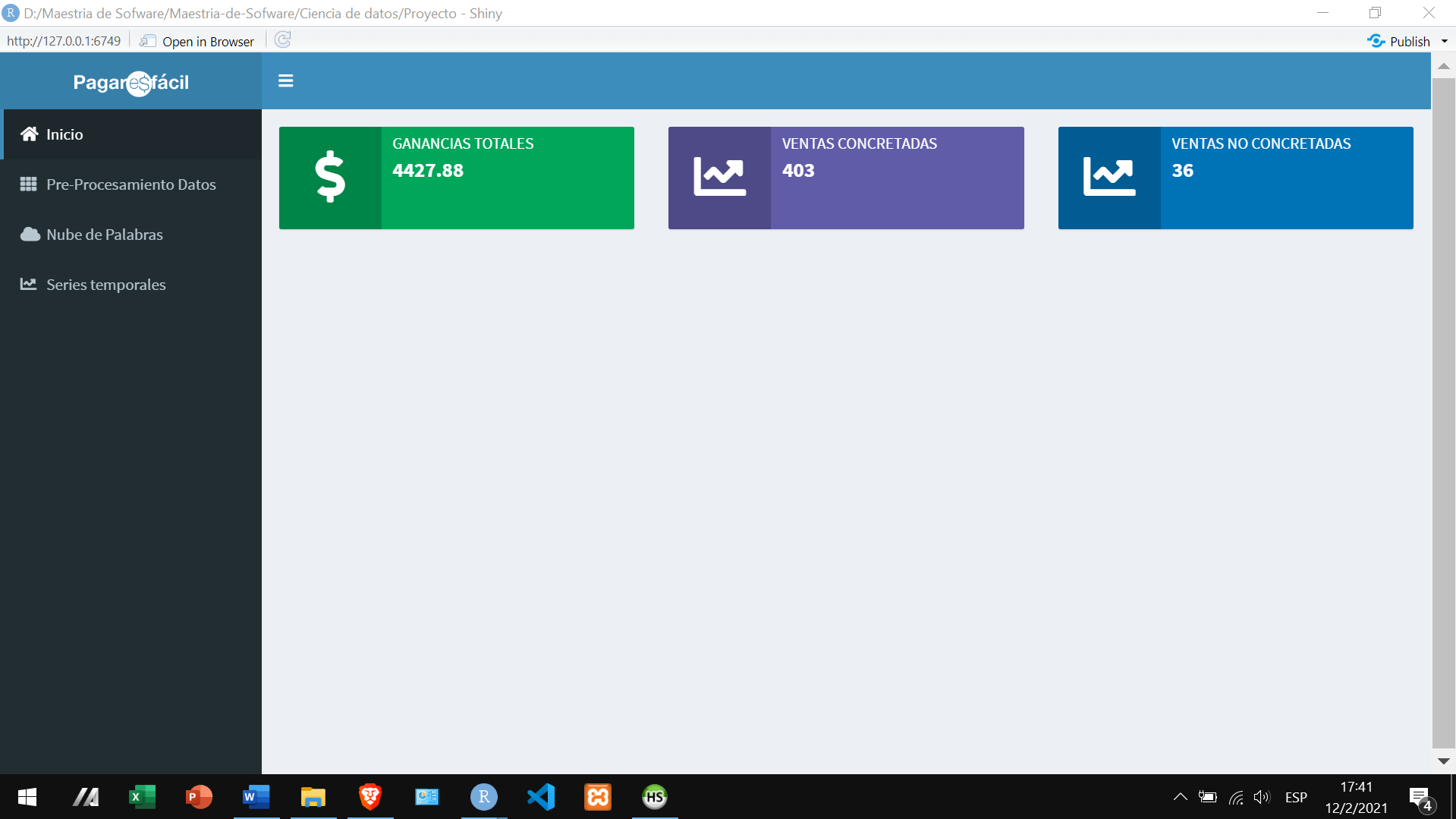


La siguiente tabla muestra las cantidades de indicadores que miden errores de pronóstico.



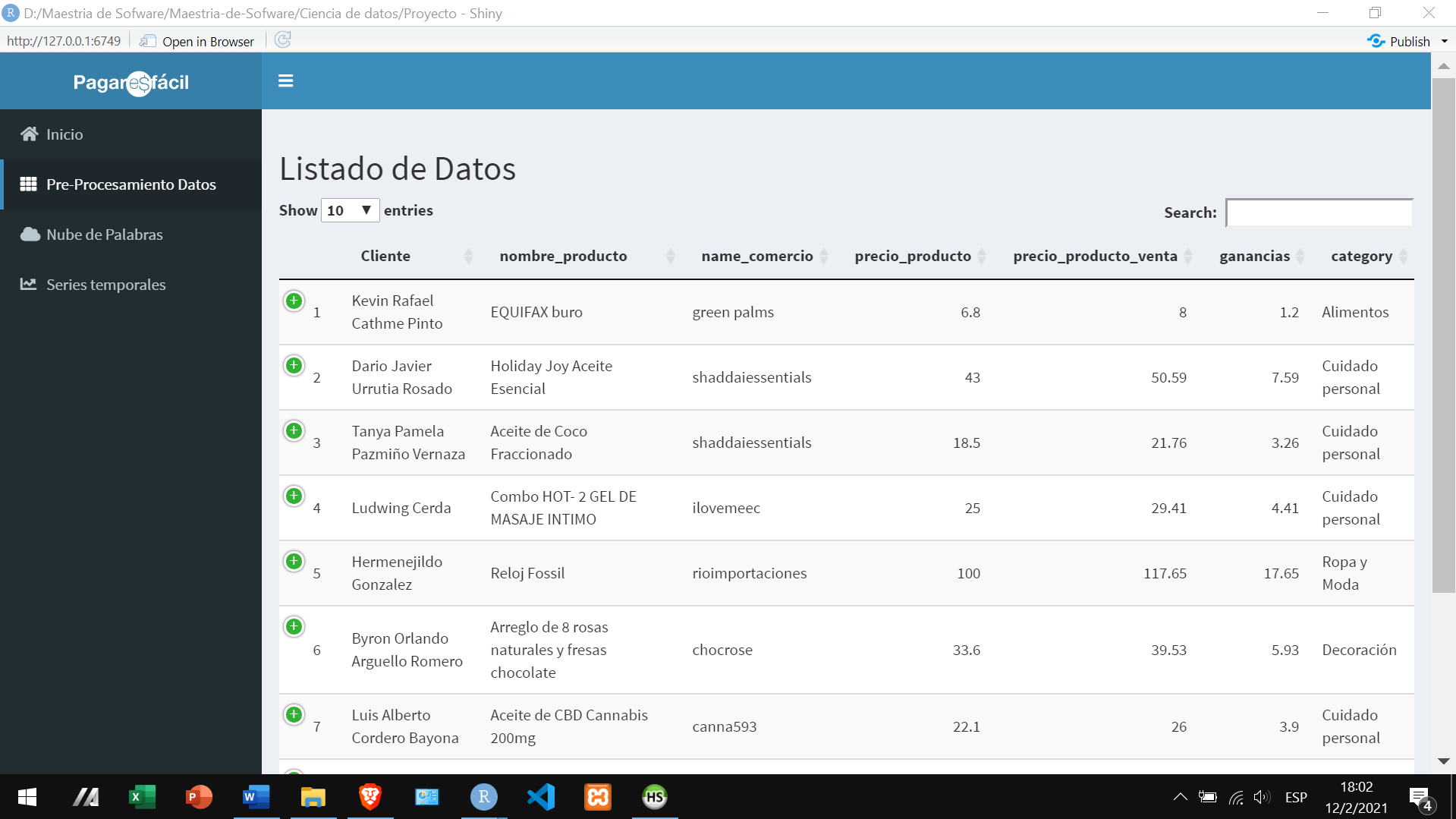
Como paso final de la metodología CRISP-DM, se muestra a continuación el despliegue del dashboard en producción.

Menú Inicio del Dashboard



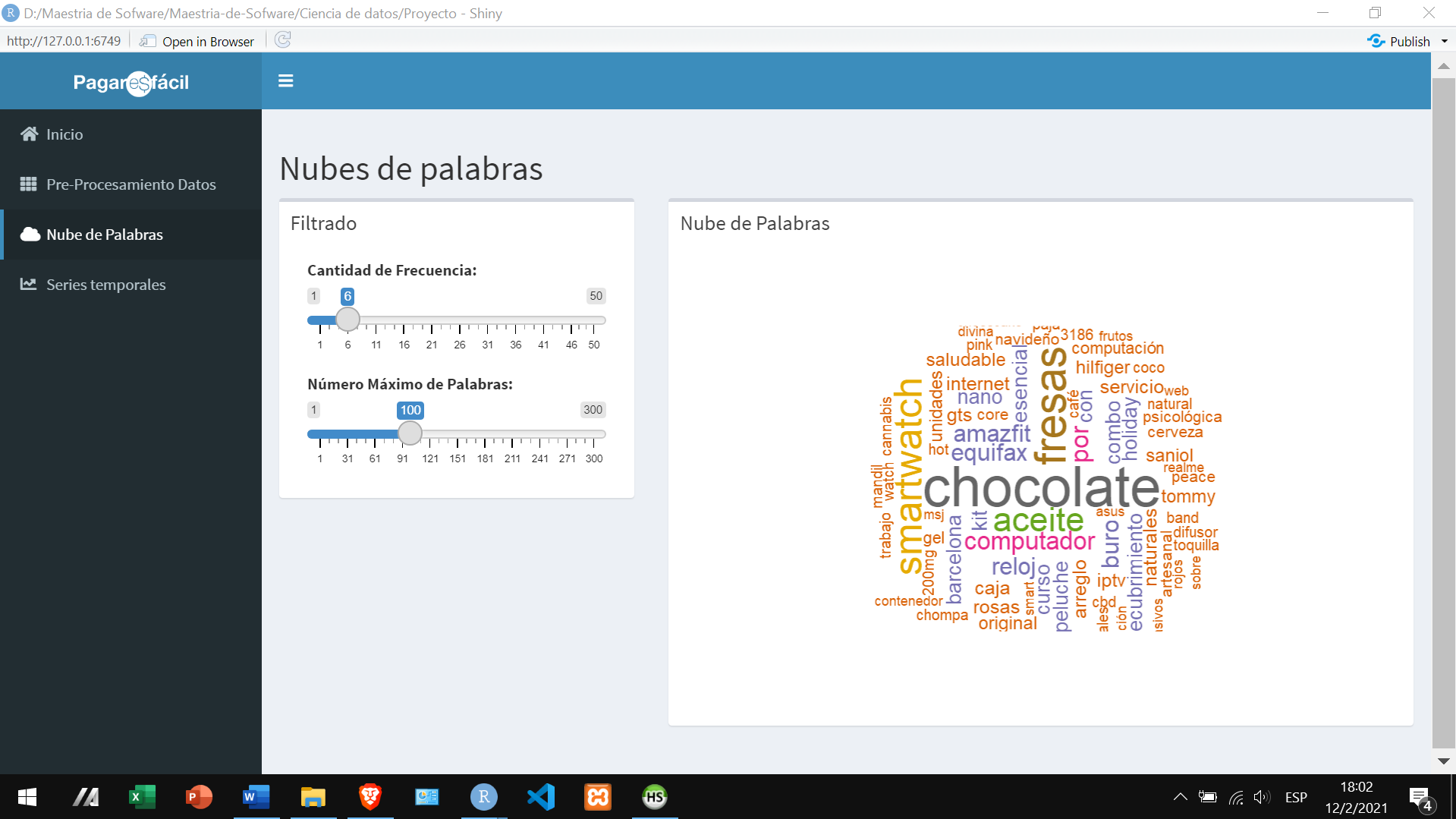
**Fuente:** Elaboración Propia

Menú de pre-procesamiento de datos.



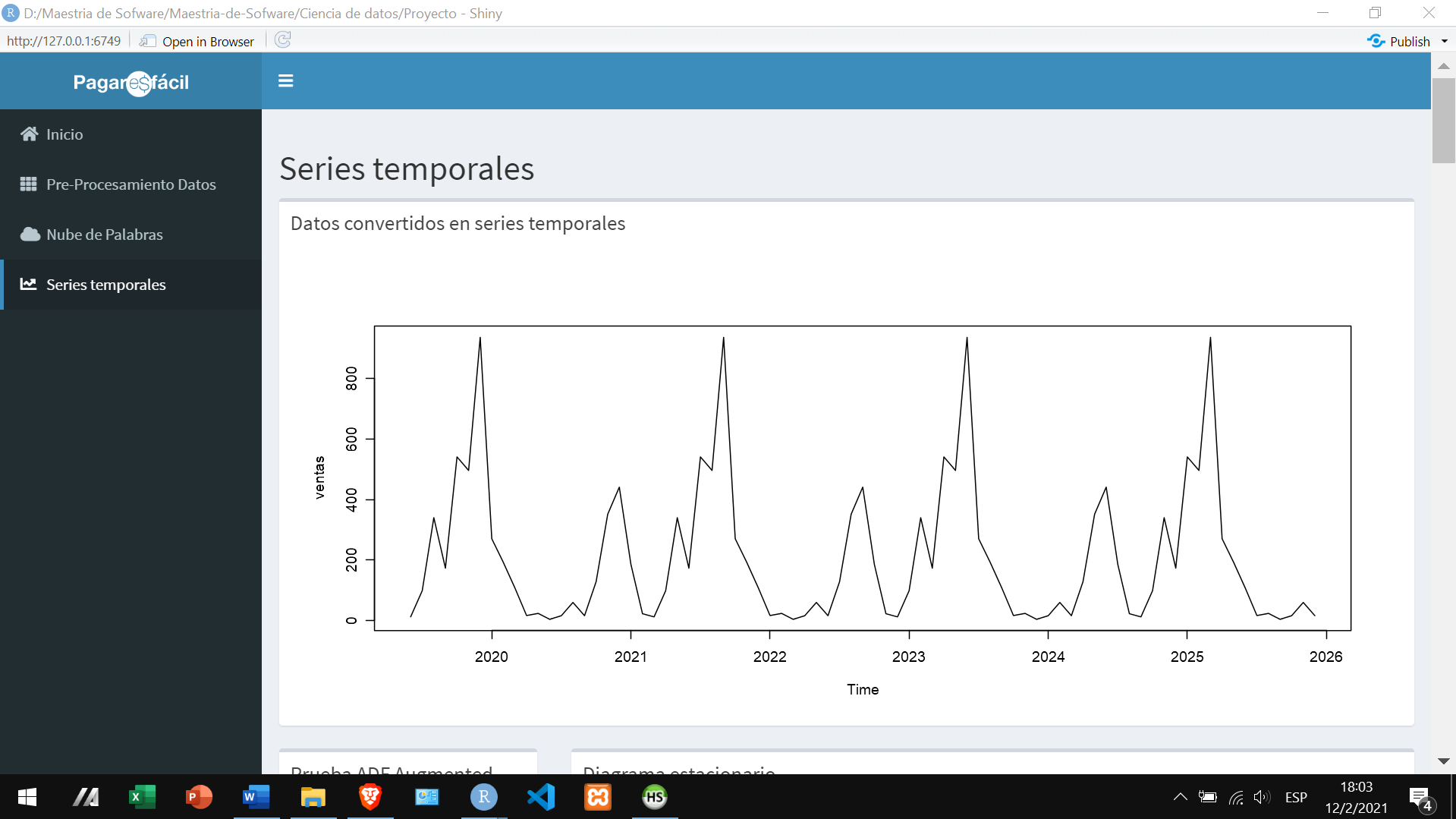
**Fuente:** Elaboración Propia

Menú de nubes de palabras



**Fuente:** Elaboración Propia

Menú de series temporales



**Fuente:** Elaboración Propia

**Conclusiones**

Esta investigación se realizó utilizando la metodología CRISP DM la cual nos permitió lograr cumplir con los objetivos planteados. En cuestión de la simulación y pronóstico de las ganancias futuras para el Marketplace de Pagar es Fácil se realizó utilizando las series de tiempo de las cuales se han evaluado los modelos más comunes en este ámbito. Mediante el análisis de los resultados se puede concluir que, para realizar un pronóstico aproximado, se debe elegir el modelo más acertado para el evento que se quiere estudiar.

El análisis estadístico permitió tomar una decisión del modelo escogido, el cual cumple con los parámetros requeridos de normalidad, varianza constante y aleatoriedad. En relación al caso de estudio presentado, el pronóstico de ganancias mensuales más altos con un intervalo de confianza del 80% entre los 12 meses del próximo año oscila entre $337.11 a $628.88 y su valor más bajo oscila entre $ -77.6 a $-252.86, ahora con respecto al intervalo de confianza del 95%, el valor más alto oscila entre $471.29 a $815.87 y el valor más bajo oscila entre $ -264.59 a $ -433.46. La simulación a partir del modelo ARIMA (3,1,6) demuestra que los pronósticos, desde un registro histórico, se adapta muy bien a los niveles máximos y mínimos, y se mantienen en el rango. Se concluye, entonces, que estos modelos no permiten simular el comportamiento exacto en el tiempo, pero es una buena herramienta con la cual se obtiene una aproximación de posibles eventos máximos y mínimos. Por último, mediante los modelos de predicción estadística y simulación es posible tener aproximaciones de los comportamientos de las ganancias para periodos cortos de tiempo, y que estas metodologías constituyen una herramienta que, una vez optimizada, permitirá obtener una aproximación de las ganancias de ventas.