**Phase 1: Exploration**

Diego Alejandro Lopez Ossa

Universidad Nacional Abierta y a Distancia – UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería – ECBTI

Especialización En Ciencia de Datos

2025

Ejercicio 1.  
para el desarrollo del ejercicio 1 tuve en cuenta alguno puntos relacionados con la limpieza de datos de los capítulos 5, 7,12 y 13 del libro [R Para Ciencia de Datos](https://es.r4ds.hadley.nz/) de de Hadley Wickham y Garrett Grolemund.

Para el primer punto del ejercicio 1 Realizare una exploración de los datos la cual la dividiré en 4 etapas, la primera etapa será en búsqueda de datos faltantes sobre cada una de las columnas que contiene mi base de datos Appendix 1 – Database.

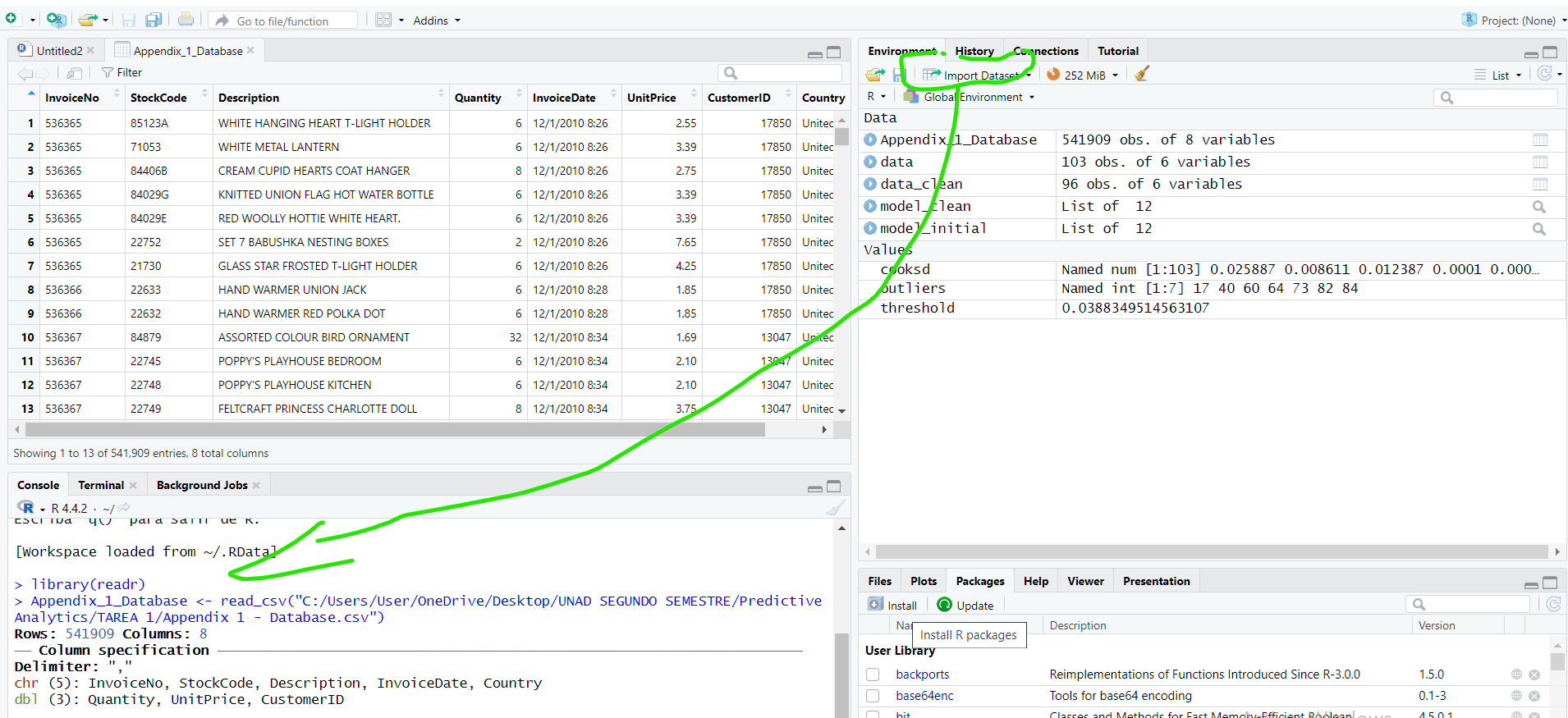


Ilustración cargue de base de datos para exploración, fuente: elaboración propia

En este primero paso por medio de la opción import dataset cargue la base de datos, el programa me muestra de forma automática que esta tiene 8 columnas de las cuales 5 tienen datos de tipo texto y 3 con datos tipo numérico, y un total de 541909 filas.

Interfaz de usuario gráfica, Texto, Aplicación

El contenido generado por IA puede ser incorrecto.

Ilustración buscando valores vacíos en columnas, fuente :elaboración propia

Por medio de la función is.na verifico si hay valores faltantes en mis columnas detectando que en las columnas description y customer id tiene 1454 y 135080 valores faltantes , si bien no tengo un contexto claro de que papel juega el customer id, lo considerare aquel valor identificador único de cada cliente que permite hacer seguimiento del comportamiento de compra, por ende es esencial para análisis que estoy llevando a cabo, creare un nuevo conjunto de datos dejando por fuera las filas donde encontré las valores nulos de la columna customer id.

La nueva base de datos se la llamara Appendix\_1\_Database\_limpio y esta contiene 406829 filas, ilustración 3.

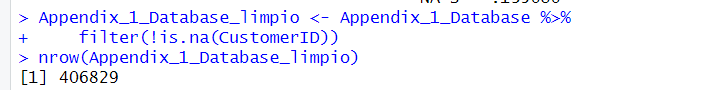


Ilustración uso de función filter para filtras datos nulos, fuete: elaboración propia.

En la segunda etapa buscare valores duplicados, si bien no me sirve como referencia ninguna de las columnas de esta base de datos, a modo de practica buscare productos similares con diferentes códigos, asumiendo que la columna StockCode debería de almacenar valores únicos para la descripción de cada producto.

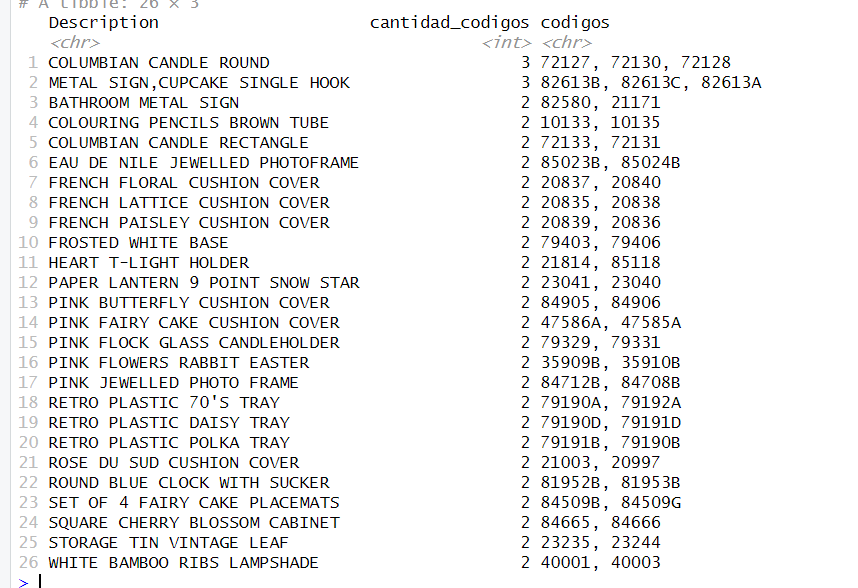


Ilustración códigos de productos por descripción, fuente: elaboración propia

En la ilustración 4 se observa que por lo menos dos códigos tienen cada descripción de producto habiendo un total de 26 productos diferentes, lo cual asumiré como un problema potencial ya que podría decir que hay inconsistencias en la codificación y posibles errores en el ingreso de datos.

En la tercera etapa verificare posibles valores ceros y negativos en las columnas Quantity y UnitPrice.

Interfaz de usuario gráfica, Texto, Aplicación

El contenido generado por IA puede ser incorrecto.

Ilustración verificación de valores negativos y ceros , fuente :elaboración propia

En la ilustración 5 tenemos en la columna Quantity 8905 valores negativos lo cual podríamos decir que se debe a devoluciones por parte de los clientes, y en cuanto a la columna UnitPrice tentemos 40 datos con valor cero lo cual podría ser muestras gratis, esto me permitiría enfocarme en los clientes con devoluciones y/o revisar el tema de campañas con muestras gratis para saber si están generando el efecto deseado en los clientes (asumiendo que no sean errores de digitación).

Para la etapa 4 realice un resumen de medidas estadísticas con el fin de encontrar patrones inusuales en los datos (outliers) obteniendo los siguientes resultados, ilustración 6.

Calendario

El contenido generado por IA puede ser incorrecto.

Ilustración resumen estadístico fuente: elaboración propia.

En cuanto a la ilustración 6 me centrare en las columnas Quantity y UnirPrice, donde en la primera se observa que la gran diferencia entre media (12.06) y mediana (5) indica una distribución sesgada hacia cantidades mayores, además los valores extremos (-80,995 y 80,995) son inusuales lo cual requiere verificación para determinarlos.

En cuanto a la columna UnirPrice hay productos con precio cero mencionado anteriormente, además se presenta un precio máximo(38,970) lo cual es extremadamente alto teniendo en cuenta que la mayoría de los productos tienen precios entre 1.25 y 3.75, la diferencia entre media y mediana sugiere precios sesgados hacia valores altos.

El segundo punto del ejercicio 1 trabajare nuevamente con la base de datos Appendix\_1\_Database, por medio de la librería tidyverse y su función mutate la cual permite agregar nuevas columnas sobrescribiendo en la base de datos existente agrego la columna de nombre precio\_total, ilustración 7.

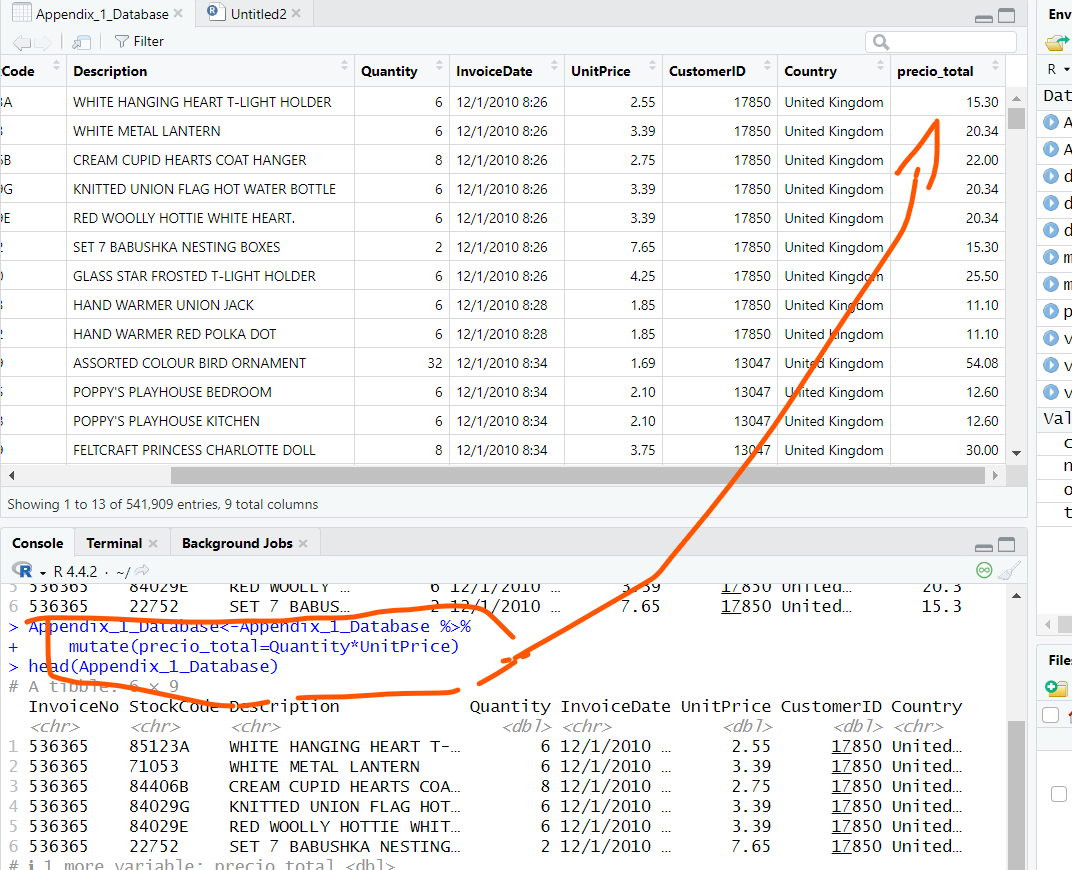


Ilustración función mutate para crear una nueva . fuente: elaboración propia.

Para el tercer punto del ejercicio uno agrupe los dato de acuerdo a las facturas que hay mediante la función groupby, después de agrupados usamos la función sum para calcular el valor total por factura, ilustración 8.

Texto

El contenido generado por IA puede ser incorrecto.

Ilustración cálculo del costo total por factura, fuente : elaboración propia.

Para el cuarto punto del ejercicio 1 por medio de la variable total\_factura, organizándola en orden descendente y con el método slice(1), selecciono el primero registro de las facturas que en este caso seria el de mayor valor., ilustración 9.

Texto, Carta

El contenido generado por IA puede ser incorrecto.

Ilustración factura con el precio más alto. fuente: elaboración propia

Para el quinto punto considero que los análisis realizados en los puntos anteriores proporcionan información crucial para optimizar las operaciones y la rentabilidad del negocio. Al calcular los precios totales por producto y por factura, la empresa puede identificar sus productos más rentables y entender mejor los patrones de compra de sus clientes. La identificación de las facturas de mayor valor y los clientes más frecuentes permite desarrollar estrategias para mejorar el servicio.

**Ejercicio 2.**  
Para el grafico del ejercicio 2 use un histograma teniendo en cuental los dicho por Wickham et al. (2023) en "R for Data Science", "los histogramas son una herramienta fundamental para entender la distribución de datos continuos, donde el ancho de los bins puede afectar significativamente nuestra interpretación de los patrones subyacentes" (p. 137).Así, por medio de la librería ggplot, si bien en un principio realice un histograma como se observa en la ilustración 10 este no me muestra un patrón que me permita tomar decisiones pues debido a valores lejanos expuestos en el punto 1 del ejercicio 1.

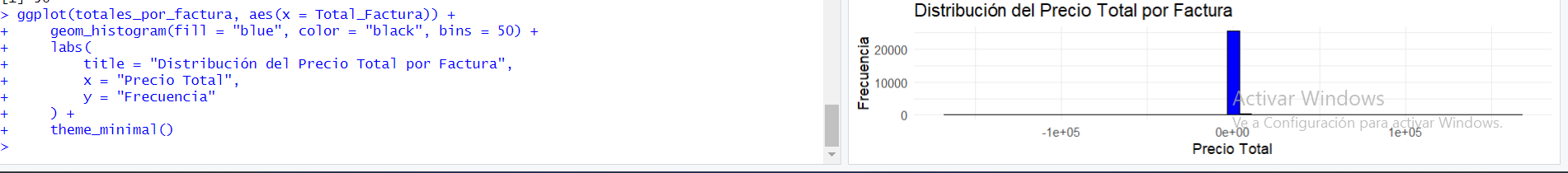
****

Ilustración histograma con calores atípicos. fuente: elaboración propia

Para solucionar este problema se ajustó el código con el fin de eliminar los valores atípicos que por lo general están por fuera del 95% de los datos, obteniendo el resultado de la ilustración 11.

Interfaz de usuario gráfica, Aplicación

El contenido generado por IA puede ser incorrecto.

Ilustración histograma con ajuste excluyendo el 5% de datos atípicos. fuente: elaboración propia.

Teniendo ajustado el grafico de la ilustración 11, este no muestra una distribución asimétrica donde la mayoría de las facturas se concentran en valores entre 0 y 500, bimodal con dos picos notable entre 100 y 400, y una disminución gradual hacia valores más altos. Considero q este patrón indica una clara segmentación natural en el comportamiento de compra de los clientes, teniendo en cuenta lo dicho por Provost y Fawcett (2013) en "Data Science for Business" donde señalan que "el análisis de patrones de compra y la segmentación de clientes basada en el valor de las transacciones son fundamentales para desarrollar estrategias de marketing efectivas y personalizadas" (p. 243).Me permitiría desarrollar estrategias de marketing más efectivas y personalizadas. Por ejemplo, se pueden crear diferentes programas de fidelización y promociones específicas para cada segmento de clientes, establecer umbrales de descuento basados en los patrones observados, y diseñar campañas dirigidas a aumentar el valor de las compras pequeñas mientras se mantiene la lealtad de los clientes de alto valor.

Conclusiones.

* La mayoría de las ventas se concentran en facturas de valor medio-bajo (entre 0 y 500), lo que sugiere una oportunidad para desarrollar estrategias que incentiven compras de mayor valor.
* El análisis de precios totales por factura y por cliente permite identificar patrones de compra y clientes clave, información vital para desarrollar programas de fidelización efectivos.
* La distribución de las ventas muestra una clara segmentación natural de clientes, lo que facilita la creación de estrategias de marketing personalizadas para diferentes grupos de compradores.

Link con repositorio para revisión de código en R: https://github.com/Dalopezos28/Phase1\_exploration.git

Referencias bibliográficas.

Wickham, H., Çetinkaya-Rundel, M., Grolemund, G. (2023). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data* (2nd ed.). O'Reilly Media.

Provost, F., & Fawcett, T. (2013*). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. O'Reilly Media.

Zero 2 Hero.(2023). *Ejemplo de un Proyecto con R y RStudio: Análisis y Limpieza de Base de Datos* [video]. YouTube. https://www.youtube.com/watch?v=Ct1CXxYe-QQ&t=622s